الحرف العربي والتقنية

أبحاث في حوسبة العربية

تأليف :

أ. مأمون صبحي الحطاب

د. أمـجد أبو جــبارة

د. إريك آتـــــــويل

أ. عمر السيد شعـبان

أ. عبدالله يحيى الفـيفي

أ. د. محمد زكي خضر

د. سامح محمد عـويضة

د. يحيى محمد الحاج

تحـــريـر:

د. يوسف سالم عيسى العريان



مباحث لغوية (12)

الحرف العربي والتقنية : أبحاث في حوسبة العربية

سلسلة (مباحث لغوية) :

يُصدِر مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية هذه السلسلة ضمن خطة عمل مقسمة إلى مراحل، تشمل مرحلتها الأولى ثلاثين عنوانا، لموضوعات علمية رأى المركز – بعد الدراسة – حاجة المكتبة اللغوية العربية إليها، أو إلى بدء النشاط البحثي فيها، ويهدف من وراء ذلك إلى تنشيط العمل في المجالات التي تُنَبِّه إليها هذه السلسلة، سواء أكان العمل علميا بحثيا، أم عمليا تنفيذيا، ويدعو المركز الباحثين كافة من أنحاء العالم إلى المساهمة في هذه السلسة.

وتودّ الأمانة العامة أن تشيد بجهد السادة المؤلفين، وجهد محرر الكتاب، على ما تفضلوا به من التزام علمي لا يستغرب من مثلهم.

والشكر والتقدير الوافر لمعالي وزير التعليم المشرف العام على المركز، الذي يحث على كل ما من شأنه تثبيت الهوية اللغوية العربية، وتمتينها، وفق رؤية استشرافية محققة لتوجيهات قيادتنا الحكيمة.

والدعوة موجّهة لجميع المختصين والمهتمين بتكثيف الجهود نحو الصعود بلغتنا العربية، وتحقيق وجودها السامي في مجالات الحياة.

المملكة العربية السعودية – الرياض

مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية

هاتف : ٠٠٩٦٦١١٢٥٨١٠٨٢ – ناسوخ ٠٠٩٦٦١١٢٥٨١٠٦٩

ص.ب : ١٢٥٠٠ الرياض ١١٤٧٣

www.kaica.org.sa



مباحث لغوية ١٢

الحرف العربي والتقنية

أبحاث في حوسبة العربية

تأليف :

أ. مأمون صبحي الحطاب

د. إريك آتويل

أ. عبدالله يحيى الفيفي

د. سامح محمد عويضة

د. أمجد أبو جبارة

أ. عمر السيد شعبان

أ. د. محمد زكي خضر

د. يحيى محمد الحاج

تحرير:

د. يوسف سالم عيسى العريان



مركز الملك عبدالله بن عبدالعزيز الدولي

لخدمة اللغة العربية

King Abdullah Bin Abdulaziz Int'l Center for

The Arabic Language

ⓒ مركز الملك عبد الله بن عبد العزيز الدولى لخدمة اللغة العربية، ١٤٣٦هـ
فهرسة مكتبة الملك فهد الوطنية أثناء النشر
مركز الملك عبد الله بن عبد العزيز الدولى لخدمة اللغة العربية
الحرف العربي والتقنية . / مركز الملك عبد الله بن عبد العزيز الدولى لخدمة اللغة العربية . - الرياض ، ١٤٣٦هـ
٢٨٨ ص ؛ ١٧ × ٢٤ سم
ردمك :٦- ٨ ـ ٩٠٦٤٨ ـ ٦٠٣ ـ ٩٧٨
١- الابجدية العربية - تاريخ   ٢- الخط العربي - تاريخ
٣- الحواسيب   أ.العنوان
ديوي ٤١١,١٠٩   ١٤٣٦/٣٩٧٠

الطبعة الأولى

١٤٣٦هـ / ٢٠١٥م

سلسلة من الإصدارات التي تعالج قضايا لغوية متنوعة

مدير المشروع :

أ. خالد بن أحمد الرفاعي

إشراف :

د.عبدالله بن صالح الوشمي


الله
بسم الله الرحمن الرحيم

كلمة المركز

يجتهـد مركـز الملـك عبدالله بـن عبدالعزيـز الـدولي لخدمة اللغـة العربية في العمـل في مجـالات متعددة تحقق تعميـق الوعي اللغوي علـى المستويات المختلفة (الاجتماعيـة والعلمية/ الأهلية والرسميـة) ؛ وذلك للسمو باللغة العربية، وترسيخ منافستهـا للغات الحضارية في العالم، وتعميق قيادتهـا الدينية والتاريخية لشعوب شتى في أنحاء المعمورة.

وامتـدادا لذلـك. ينشط المركز في مجـال النشر، مستقطبـاً الأعمال العلمية الجـادة وفق لائحـة معتمدة منظّمة لذلك، كما ينشط في مجـال التأليف من خلال استكتاب مجموعة كبيرة من الباحثين ؛ لتأليف عدد متنوع من الإصدارات النوعية المقـروءة التي تعالـج عنوانـاتٍ يقتنصها المركز، ويلفـت الانتباه إليهـا، ويعلن من خلالهـا الفرص الممكنة لخدمـة اللغة العربيـة في المجالات المختلفـة، ملبّياً بذلك الحاجـات التـي يلمس المركز تطلّع المكتبة اللغوية العربيـة إليها، ولافتاً الأنظار إلى أهميـة التعمق فيهـا بحثياً، واستكشاف مـا يمكن عمله تنفيذيـاً في هذه المجالات. ويسعـد المركز بـأن استقطب في المرحلة الأولى من هذا المشـروع ما يربو على مئتي باحـث، موسِّعاً دائـرة المشاركة محليـاً وخليجيا وعربيا وإسلاميـا وعالميا، ومنوّعاً مسـارات البحـث الرئيسية والفرعية، ومنفتحـا على كل ما من شأنـه خدمة اللغة العربية بجميع الوسائل والأطر.

ويمثّـل هـذا الكتـاب واحدا من الكتـب التي صـدرت ضمن سلسلـة (مباحث لغويـة) يحتـوي عددا من الأبحـاث لأساتذة مرموقين؛ استجابوا لمـا رآه المركز من الحاجة إلى التأليف تحت هذا العنوان، وبادروا إلى ذلك مشكورين.

وتـودّ الأمانـة العامـة أن تشيد بجهد السـادة المؤلفين، وجهد محـرر الكتاب، ومديـر هـذا المشروع العلمي علـى ما تفضلوا به مـن التزام علمـي لا يستغرب من مثلهـم، وقـد ترك المركـز للمحرر مساحة واسعـة من الحرية في اختيـار الباحثين ووضـع الخطة العلمية - بالتشـاور مع المركز -؛ سعيـاً إلى تحقيق أقصى ما يمكن تحقيقه من الإفادة العلمية، مع الأخذ بالاعتبار أن الآراء الواردة في البحوث لا تمثل رأي المركـز بالضرورة، ولكنها من جملـة الآراء العلمية التي يسعد المركز بإتاحتها للمجتمـع العلمـي وللمعنيين بالشأن اللغـوي لتداول الرأي ، وتعميـق النظر، ونلفت انتبـاه القـارئ الكـريم إلى أن ترتيب أسمـاء المؤلفين على الغـلاف موافق لترتيب أبحاثهـم في الكتاب، وهي خاضعة للرؤية المنهجية التـي تفضّل المحرر - مشكورا- باقتراح خطتها.

والشكـر والتقديـر الوافـر لمعـالي وزيـر التعليم المشـرف العام علـى المركز، الـذي يحث على كل ما من شأنه تثبيت الهوية اللغويـة العربية، وتمتينها، وفق رؤية استشرافية محققة لتوجيهات قيادتنا الحكيمة، ويمتد الشكر لمعالي نائبه، وللسادة أعضاء مجلس الأمناء نظير الدعم والتسديد لأعمال المركز.

والدعـوة موجّهة لجميـع المختصين والمهتمـين بتكثيف الجهود نحـو النهوض بلغتنا العربية، وتحقيق وجودها السامي في مجالات الحياة.

المقدمة

إنّ الحمــد لله، نحمــده، ونستعينه، ونستغفره، ونعوذ بــالله من شرور أنفسنا، ومــن سيئات أعمالنا، من يهده الله فلا مضــلّ له، ومن يضلل فلا هادي له، وأشهد أنّ لا إله إلاّ الله، وحده لا شريك له، وأشهد أنّ محمدًا عبده ورسوله؛ أما بعد،،

فقــد أيد الله -تبارك وتعالى- رسله بآيات، ثم اصطفى لخاتمهم منها كلامه العربــي الخالــد المبين؛ فكانت آيته من جنس البيان (الرَّحْمَــنُ . عَلَّمَ الْقُرْآنَ . خَلَقَ الإِنْسَانَ . عَلَّمَهُ الْبَيَانَ) وكانت (بِلِسَانٍ عَرَبِيٍّ مُبِينٍ). وفي هذا ما يكفي للإشارة إلى فضل البيان عموما واللسان العربي خصوصا.

ولعلــوم الحاسب فضلها المستمد من كــون الحاسب آلة تمد وتقوي ما ميز الله بــه بني آدم من عقل؛ كما أن المقراب امتداد لقوة الإبصار. لذا، فيمكننا القول بأن نسبــة رقي الحاسبات علــى غيرها من الآلات كنسبة رقي العقل على الجسد. ثم إن برمجة هذه الآلات -التي ابتدأت ببعض أعمال الإخوة بني موسى- تحتاج إلى لغات خاصة بها، مما يعود عليها بما أسلفنا من فضل البيان واللسان.

يتنــاول كتابنــا هذا الحاســب في خدمــة اللغة العربيــة. وهو في هيكلــه يشبه الشجــرة: إذ يبدأ بثلاثة بحــوث في البنى التحتية: أحدها يناقش الأساس التقليدي لبرامــج حوسبة اللغــة، وهو تمثيل اللغة العربية في الحاســوب، والآخران للمدونات اللغويــة وقواعد البيانات التي تنبنــي عليها النظم المتعلّمة. ثم ينتقل إلى ما يسامي الجــذع -وهي الأدوات التــي تحمل غيرها مــن التطبيقات المثمــرة- فجاء ببحثين استقصــى أحدهمــا التقنيــات التقليديــة والإحصائيــة لمعالجة اللغــات الطبيعية وتخصصــص الآخر في طريقــة لتعلم التشكيل آليا. ثم خُتــم الكتاب بفرعين مثمرين

استقصيا تطبيقات الترجمة الآلية من العربية وإليها وتطبيقات التعرف الآلي على قراءة القرآن الكريم.

وكان حرصنا منصبا في هذا الكتاب العربي على المقدمات الاستقصائية لأنه باكورة في سلسلة، فناسبه التمهيد لما بعده. كما حرصنا أن يكتب موضوعاته نخبة من أساتذة الجامعات ومدراء الشركات المتخصصة في المملكة العربية السعودية والمملكة الأردنية الهاشمية والمملكة المتحدة والولايات المتحدة الأمريكية. وإننا لنرجو أن تستمر هذه السلسلة المباركة وأن يَفيد من هذا الكتاب وما بعده باحثو اللسانيات وباحثو الحاسوب وغيرهم، والله ولي ذلك والمبارك فيه.

المحرر :

د.يوسف بن سالم العريان

حول نظام تمثيل الحرف العربي

أ. مأمون صبحي الحطاب (*)
m.hattab@arabtext.ws

(*) دار حوسبة النص العربي
باحث في حوسبة اللغة العربية، قام بتطوير عدد من الأنظمة المحوسبة في مجالات: الصرف والنحو، ومحركات البحث العربية، وتطويع التقنية لخدمة ذوي الاحتياجات الخاصة العرب.
وهو عضو مؤسس لجمعية حوسبة اللغة العربية وإثراء المحتوى العربي على شبكة الإنترنت.

١. المقدمة

تعتمـد برامـج التحليـل اللغوي علـى بنية الكلمـة ورسمها كمصـدر لمعلومات التحليـل اللغـوي. فعمل الحاسوب يتكون من ثلاث مراحـل هي مرحلة الإدخال، ثم مرحلـة المعالجة، تتبعهما مرحلة إخراج النتائج. ولم تحقـق برامج التحليل اللغوي العربيـة -حتى الآن- نتائج تـؤدي إلى بناء تطبيقات لغوية لها أثر مرض في الواقع. ويعرض البحـث خصائص نظم تمثيل العربية وبعضـ مشاكلها في سياق مقارنتها بنظم تمثيل اللغة الإنجليزية، وهي اللغة العالمية الحالية.

لقـد أدت اللغة العربية دورا بارزا في نقل الحضارة الإنسانية واستيعابها لعدة قـرون دونمـا انقطاع. ويعبر عن هذا الـدور الحضاري مؤسس علـم اللغة الحديث إدوارد سابير Edward Sapir ١٩٤: ١٩٢١، حيث يقول:

"إن اللغات الخمس التي كان لها دور رئيس في حمل الحضارة والإنسانية هي: الصينية القديمة والسنسيكريتية والعربية واليونانية واللاتينية".

والملاحـظ أن اللغـة العربية هي الوحيدة من بين هـذه اللغات الخمس التي ما زالـت تؤدي دورهـا دون انقطاع. وقد شهد للعربية بأنها لغـة علمية- العالم اللغوي رونالـد لانغيكـر Ronald Langacker (١٨٢: ١٩٧٣) الذي يرى أن طبيعة الكلمات المستعـارة من لغة مـا تعكس مدى تأثيرها في اللغة المستعـيرة، وأن نسبة كبيرة من الكلمات العربيـة الداخلة في الإنجليزية هي كلمات علمية مثل Zero وAlchemy و Algebra وغيرها كثير. وهي كلمات دخلت الإنجليزية عن طريق الإسبانية. ويذكر أن الدكتـور وجيـه حمد عبد الرحمـن قد أثبت وجود ١٠٠٠ جـذر عربي في معجم أوكسفورد الإنجليزي (١).

أما المستشرقة الألمانية ريجارد هونيكه فتقول في كتابها الشهير "شمس العرب تسطع على الغرب" (٢٤٣: ١٩٦٣) : إنه قبل ستمئة عام كان لكلية الطب الباريسية أصغر مكتبة في العالم، لا تحتوي إلا على مؤلف واحد، لعربي عظيم . وكان هذا الأثر العظيم هو المرجع الأساسي لمدة تزيد عن أربعمئة عام بعد ذلك التاريخ دون أن يزاحمه مزاحم، أو تؤثر فيه أو في مكانته مخطوطة من المخطوطات الهزيلة التي دأب في صياغتها كهنة الأديرة قاطبة. وهذا العمل الجبار خطته يد الرازي (أبو بكر محمد بن زكريا) . وقد اعترف الباريسيون بقيمة هذا الكنز العظيم وبفضل صاحبه عليهم وعلى الطب إجمالا فأقاموا له نصبا في باحة القاعة الكبيرة في مدرسة الطب لديهم، وعلقوا صورته في شارع سان جيرمان.

بالرغم مما تقدم من حقائق علمية ثابتة فقد تعرضت اللغة العربية لحملات مغرضة تفتقر إلى الموضوعية وتتناقض مع حقائق علم اللغة الحديث. وقد انصب الهجوم على عدة محاور منها:

١. الادعاء بعدم صلاحية الخط العربي والدعوة لتلتينه.
٢. الادعاء بأن العربية لغة دينية - بالمفهوم الكهنوتي للدين — وليست لغة علمية.
٣. الادعاء بجمود العربية الفصحى وضرورة السماح بالتغيير اللغوي المتمثل في العاميات والدعوة لإحلالها محل الفصحى.
٤. القول بافتقار العربية للعدد الكافي من اللواصق (السوابق واللواحق) لترجمة ما يعادلها في اللغات الأوروبية الحديثة التي تنقل العربية عنها في القرن العشرين.

٢. تاريخ تطور نظام كتابة العربية

عرفت العرب الكتابة في جاهليتها —قبل الإسلام- واعتبرتها شرطا في كمال الرجل العربي، وتعود معرفتهم بالكتابة إلى اتصالهم بالأمم المتحضرة في بلاد

اليمن وتخـوم الشام، فأنشأ الأنباط –مثلا- ممالكهم على أطراف بلاد الشام في الناحيـة الشمالية الغربية من شبه الجزيرة العربية (١٦٩ق م – ١٠٦م) ،واتخذت البـتراء "سلع" عاصمة لهـا، وكانت لهم صلات بالأراميـين، فتأثروا بهم وتحدثوا لغتهم، واستنبطوا لأنفسهم خطا خاصا بهم عرف بالخط النبطي، اشتق منه عرب الشمال الخط الأنباري والخط الحيري، أو الخط المدور والخط المثلث.

ونجد عند العودة إلى تاريخ الكتابة العربية أنها مرت بثلاث مراحل (٢) هي: مرحلـة النشأة التي تمتد حتى ظهـور الإسلام، ومرحلة الاستخـدام الواسع وتمتد حتى بعيد منتصف القرن الثاني الهجري ثم مرحلة تقعيد الإملاء.

فقـد شجـع الإسلام على تعلم الكتابـة، وسلك في ذلك وسائـل مختلفة، حتى أن رسـول الله صلى الله عليه وسلـم اشترط لفكاك الأسير من قريش في بدر تعليم عشرة من صبيان المدينة الكتابة، فراجت الكتابة في عصره صلى الله عليه وسلم، حتى بلغ عدد كتاب الوحي أكثر من أربعين كاتبا.

وتعد الحجاز أول بلاد العرب معرفة للكتابة، وكانت قريش في مكة، وثقيف في الطائـف أكثر القبائل شهرة بها، ومن أبنائها اختير كتاب صحف أبي بكر الصديق رضـي الله عنـه، وكان عمر بـن الخطاب رضي الله عنـه يقول كمـا روى جابر بن سمـرة: لا يملين في مصاحفنا هذه إلا غلمـان ثقيف. وعندما جمع عثمان بن عفان رضي الله عنه مصاحفه قال: اجعلوا المملي من هذيل، والكاتب من ثقيف. (٢)

ثم أتت مرحلة التقعيد، وهي المرحلة التي تأثرت بالقواعد الصرفية والنحوية، مما جعـل الإملاء العربي، في مواضع كثيرة، يشـير إلى قواعد النحويين أكثر مما يصف واقعا كتابيا ملموسا. (٣) حتى رأى بعض الباحثين المعاصرين (٥) أن اللغة العربيـة مكونة من ثلاثـة أنظمة، هي: النظام الصوتي والنظـام الصرفي والنظام النحـوي، وقائمة من الكلمات التي لا تنتظم في جهاز واحد. وأن هذه الأنظمة تكون

معينا صامتا، فإذا أردنا أن نتكلم أو أن نكتب؛ نظرنا في هذا المعين الصامت فوضعنا محتوياته في حالة عمل وحركة، فأخذنا منه الكلمات ورصفناها على شروط الأنظمة، أي بحسب قواعد اللغة.

كان الإملائيون العرب أنفسهم نحويين وصرفيين، لذلك لم يفصلوا -عند وضعهم لقواعد الإملاء- الكتابة عن اللغة، ولم يفصلوا قواعد الإملاء عن قواعد اللغة، بل إنهم أقحموا قواعد اللغة في الكتابة والإملاء، مما جعل الإملاء العربي -في مواضع كثيرة- يشير إلى قواعد النحويين أكثر مما يصف واقعا كتابيا ملموسا. (٥)

٣. تمثيل الكتابة لصوت الكلمة

حاول الرسم الكتابي منذ نشأته أن يجاري المنطوق في وظيفته وأن يكون مطابقا لأصوات الكلام، فتطور في سبيل ذلك من الصوري إلى المقطعي ثم إلى الأبجدي، واتخذ رموزًا مميزة للحروف والحركات، وابتدع علامات خاصة لتمثيل الظواهر الصوتية الأخرى وتقريب المكتوب من المنطوق. ومما ابتدعته العربية في هذا الشأن الشدة والمدة وعلامات الحركات والسكون والنقط الذي يختلف عددًا ومكانًا وأدخل – لما شاع اللحن إثر دخول غير العرب في الإسلام- على الكتابة العربية ما يطورها مثل:

أ. النقط

وتجمع معظم المصادر على أن أول من أدخل النقط هو عالم اللغة المشهور أبو الأسود الدؤلي (توفي عام ٦٨٨م) وذلك لما لوحظ من لحن في تلاوة كتاب الله العزيز.

ب. الإعجام

ويعرّف ابن جني الإعجام بأنه إزالة العجمة. وهو مشتق من "أعجم" أي أزال

العجمة ومنها كلمة معجم الذي يزيل عجمة الكلمات بإيضاح معناها.

ت. الشكل

وأول مـن أدخل الحركات القصـيرة الضمة والفتحة والكسـرة هو الخليل بن أحمـد الفراهيدي (المتـوفي عام ٧٨٦م). كما أدخل السكـون (ْ) الذي يعني غياب الحركة.

وبإدخـال هـذه الاصلاحات الضرورية علـى نظام كتابة العربيـة أصبح ذلك النظـام أقرب ما يكون إلى نظـام الكتابة الصوتـي (Phonetic Alphabet) الذي يحقق تناظرا بين الصوت والحرف وذلك على النحو الآتي:

أولا : التناظر بين الحروف الصائتة والأصوات الصائتة (ويلاحظ هنا أن كل حرف يمثل صوتا واحدا لا غير والعكس صحيح):

/ب	/b	باب	/bæb/
/ت	/t	تاب	/tæb/
/ث	/ʔ	أثاث	/ʔaθæθ/
/ج	/ʤ	جار	/ʤær/
/ح	/ħ	حَمَلَ	/ħamal/
/خ	/x	خلود	/xulu:d/
/د	/d	دار	/dær/
/ذ	/ð	ذاك	/ðæk/
/ر	/r	رب	/rabb/
/ز	/z	زار	/zær/
/س	/ʔ	أساس	/ʔasæs/
/ش	/ʃ	شمس	/ʃams/
/ص	/S	صَفّ	/Ṣaff/
/ض	/D	ضباب	/Ḍabæb/
/ط	/T	طبّ	/Tibb/

/ظ	/Z	ظُلم	/Zulm/
/ع	/ʕ	علم	/ʕilm/
/غ	/ɤ	غَيم	/ɤaim/
/ف	/f	فشل	/faʃal/
/ق	/q	قَلب	/qalb/
/ك	/k	كتاب	/kitæb/
/ل	/l	ليل	/lail/
/م	/m	مَلَك	/malak/
/ن	/n	نُور	/nu:r/
/ه	/h	هَرَبَ	/haraba/
S/	/ʔ	أب	/ʔabb/

ثانيا : الحروف الصائتة والحركات

/و	/:w,u	وراء	/ʔ:wara/
		نور	/nu:r/
/ي	/:j,i	يوم	/jaum/
		فيل	/fi:l/
/ا	/æ	سماء	/ʔsamæ/
/ا	/:a	غار	/ɤa:r/
/ـَ	/a	كَتَبَ	/kataba/
/ـِ	/i	كُتِبَ	/kutiba/
/ـُ	/u	كُتب	/kutub/

وهكـذا يتضـح أن درجة كفاءة الكتابة العربية تضاهـي أحدث وأدق ما توصل إليـه علماء اللغة والأصـوات في الأوساط الغربية ألا وهو "نظـام الكتابة الصوتية" الذي وضـع للتغلب على الصعوبات المستعصية التـي تواجهها أنظمة الكتابة لمعظم اللغات الحديثة التي تعرضت لحدوث بون شاسع بين نظامي الكتابة والنطق.

في المقابـل نجـد اللغة الإنجليزية أبعـد كثيرا عن نظام الكتابـة الأمثل وذلك للأسباب الآتية:

١- تمثيـل بعض الحروف لأكثر من صـوت. ففي اللغة الإنجليزية -وهي اللغة العالمية الأولى في عالمنا المعاصر- نجد ما يأتي:

أ. يرمز الحرف "a" إلى الأصوات التالية : \æ\ ، \c:\ ، \a:\ ، \i\ ، \i:\ كما في الكلمات fat ، ball ، father ، village و Caesar على التوالي.

ب. يرمز الحـرف "s" إلى \z\ ، \s\ ، \iz\ ، \3\ ، \ʃ\ ، \ф\ كما في الكلمات bags ، cats ، dishes ، measure ، pressure و island على التوالي.

وينطبق هذا على كثير من الصوامت والصوائت الإنجليزية.

٢- تمثل بعض الأصوات بأكثر من حرف واحد. فالصوت \i:\ وهو صوت مد في الإنجليزية يمثل على النحو التالي وبالحروف التالية :

ee، ea ، ei ، ie ، ey- ، e ، eo ، æ

وذلـك كمـا في الكلمـات see ، tea ، receive ، thief ، key ، crises ، people ، Caesar على التوالي.

٣- توجـد حروف صامتة في معظم الكلمات وهـو ما يضاعف صعوبة القراءة والكتابـة في الإنجليزيـة للناطقـين بها من ابنائهـا وغيرهم. وفيما يلـي أمثلة على ذلك:

المثال	الصامت
Thumb	B
Wednesday	D
Ate	E
Psychology	P
Apostle	T
River	r.
Damn	N
Walk	L
Wrong	W
Knife	K, E
Sign	G
Honest	H

وقد نجم عن هـذه العوامل وأمثالها بون شاسع بين نظامي الكتابة والنطق في لغة العلم والحضارة في يومنا هذا مما حدا بالبرلمان البريطاني لطرح هذه المسالة المؤرقة على بساط البحث لإعادة التوازن بين النظامين.

إن الذي أدى إلى حدوث مثل هذا الخلل الجسيم في النظام اللغوي الإنجليزي هـو تغير النطـق دون أن يواكبه تغير في نظـام الكتابة. حتـى إن الكاتب الإيرلندي الشهير جـورج برناردشو George Bernard Shaw يوضـح حجم المفارقة إذ يقول إن كلمة "ghoti" يمكن أن تنطق "fish" إن استعضنا عن حروف الأولى بالأصوات المأخوذة من الكلمات أدناه:

laugh = f \laif\

women = i \wimin\

nation = ∫ \nei□ạn\

هذا القصور وما يمثله من تحد حاسوبي لا يقارن بذلك الموجود في نظام كتابة اللغة العربية، ومع ذلك فإن التقدم في نظم معالجة اللغة الإنجليزية بالحاسوب لم يتوقف طويلا أمام تلك التحديات.

٤. بعض مشاكل تمثيل الكتابة العربية

كان الإملائيون العرب أنفسهم نحويين وصرفيين -كما ذكرنا سابقا- فلم يفصلوا، عند وضعهم لقواعد الإملاء، الكتابة عن اللغة، ولم يفصلوا قواعد الإملاء عن قواعد اللغة، بل أقحموا قواعد اللغة في الكتابة والإملاء، مما جعل الإملاء العربي، في مواضع كثيرة، يشير إلى قواعد النحويين أكثر مما يصف واقعا كتابيا ملموسا. (٥) وجاء هذا التقعيد على حساب دقة التمثيل الصوتي للمنطوق وأدى إلى قصور فيه، كما جعل قواعد الإملاء عرضة لاختلاف وجهات نظر العلماء تبعا لاختلاف وجهات نظرهم في الصرف والنحو.

ولم يكن واقع التطبيق بهذه المعيارية الصارمة، فاعترى الرسم الكتابي العربي مشكلات متعددة، وأخطاء شائعة عند الاستعمال. فمن المشكلات البارزة غياب الحركات وعدم اندغامها في بنية الكلمة العربية المكتوبة. وهي تزود من الذاكرة بناء على الخبرة اللغوية المستكنَّة في الذهن، وعلى السياق اللغوي الذي يحصر المعنى. وتُلَفَظ الكلمة في النصوص غير المشكولة كما تفهم من السياق وكما يحددها الحس اللغوي مما يصعب للحاسوب معها.

إن النظام الكتابي الحالي للغة العربية هو نظام صامتي Consonantal عمومًا لا يعترف إلا بالحروف الصحيحة وحروف المد واللين، ولا يعترف بالحركات

حروفًا. فكل ما ليس له رمز في الألفباء العربية لا يعد حرفًا، على الرغم من قول علماء العربية (٣) إن الحركات أبعاض حروف المد واللين، وهي تماثلها من حيث النوع ولكن تختلف عنها من حيث الكم، و إدراكهم قيمة الحركات في التمييز بين الألفاظ المعجمية والصيغ الصرفية والحالات الإعرابية.

وقد استعيض عن دمجها في بنية الكلمة المكتوبة بوضع علامات لها Diacritics تكتب خارج الكلمة فوق الحروف أو تحتها.

ومن المشكلات رسم الهمزة الذي يختلف حسب موضعها من الكلمة (أي مجيئها أولًا أو وسطًا أو آخرًا) وحركة ما قبلها أو سكونه، وحركتها هي أو سكونها. ويتصل بهذا الجانب التباس همزة القطع بهمزة الوصل التي يُؤتى بها توسلًا إلى النطق بالساكن في أول المجموعة الكلامية Utterance، وهي تختفي إذا سُبِق الساكن بحركة في درج الكلام. ومع ذلك نظل نحتفظ برسمها في ذلك الموضع، تنبيهًا على أن هذا الموضع هو مما يستوجب إدخال الهمزة عندما لا تكون هناك حركة. وواضح ما تدخله همزة الوصل من تغيير في النمط المقطعي للكلمة.

ومن المشكلات كذلك كتابة الألف في نهاية الكلمات من أفعال وأسماء قائمةً أو غير قائمة (أي مقصورة في الاصطلاح الشائع)، والتفريق بين هاء الضمير (كما في له وعنده) والتاء المربوطة (كما في طالبة) إذ يحدث اللبس فيهما عندما لا توضع نقطتان للتاء المربوطة. و كذلك التفريق بين التاء المربوطة التي يوقف عليها بالهاء (مثل فاطمة) والتاء المفتوحة التي يوقف عليها بالتاء (كما في قالت وأوقات).

وقد يُعَد من المشكلات وصل الحروف بعضها ببعض، لأنه يجعل من الكلمة وحدة إملائية لا وحدة معجمية أو صرفية بحتة، ويصعب تحليل كلمات مثل (سألتكه) إلى عناصرها الأولى من صرفية ومعجمية.

ومـن المشكلات أن الواو تمثـل حرفين هما واو المد وواو اللـين، وأن الياء تمثل حرفـين كذلك هما ياء المد وياء اللين، مع أن اللين ذو طبيعة صامتة والمد ذو طبيعة صائتة. (٧)

ومن المشكلات زيادة حروف لا تنطق، كزيادة ألف بعد واو الجماعة التي يُسند الفعـل الماضـي والمضارع المجزوم أو المنصوب إليها (دَعَـوْا، لم يَدْعُوا، لن تدعوا). وربمـا كانت هذه الزيـادة ذات فائدة في التمييز بين واو الجماعـة التي أسند إليها الفعـل والواو الأصلية التي ينتهي بها الفعل الناقص غير المسند، مثل يدعو ويرجو. إلا أنـه صـار مـن الأخطاء الشائعـة إدخال هذه الألـف بعد واو الفعـل غير المسند (يدعو)، وبعد جمع المذكر السالم المرفوع المضاف (مهندسوا المشروع).

وعكـس ذلك نقص بعض الحروف كنقص الألف في أسماء الإشارة (مثل هذا وهذه). وقد أصبحت هذه الكلمات وأشباهها بحكم الشيوع والإلف بها شيئًا معتادًا لا يثـير في نفس العربي أية مشكلة، خلا بعض الصعوبات في مراحل التعليم الأولى. وربما كانـت هذه الكلمات وأضرابها مشكلة لدى المتعلمـين من الأجانب. وكاختفاء لام التعريف -نطقا- قبل الحروف الشمسية ووهو من قبيل الإدغام□□□.

ومن الأخطاء الشائعـة وضع ياء بعد ضمير المخاطبة (قرأتي، لكي)، والخطأ في كتابة "ابن" بين علمين، والطريقة التي تكتب بها بعض الألفاظ مثل مئة بالألف، و"داوود" و"طاووس" بـواو واحدة، والتـي والذي واللذين بلامـين فتلتبس بلفظي المثنى اللذين واللتين.

والخطـأ في كتابـة الاسم المنقوص، ومسائل الفصل والوصـل في أمثال: لئن، ولئلا وطالما وإنما وثلاثمئة.

□□□ يُنظـر في مشـكلات الرسـم والأخطاء الشائعة الكتـب الآتية: الإمـلاء العربي لأحمد قبش، والإملاء والترقيم في الكتابة العربية لعبد العليم إبراهيم ، وقواعد الإملاء لعبد السلام هارون

ويمكن التفريق بين الواو المدية والواو اللينة، والياء المدية، والياء اللينة، بوضع ضمة قبل واو المد، وكسرة قبل ياء المد، كما فعل القدماء.

إن المشكلات السابقة ليست بالمستعصية، ويمكن للبرامج الحاسوبية أن تقوم بمعالجة كثير منها، وقد أحرز نجاح لا يستهان به في هذا المجال. وبقي بعضها عقبة أمام برامج التحليل اللغوي.

كما يجدر التنبه هنا إلى أن علامات الترقيم تساعد كالضبط تمامًا على الفهم السليم للنصوص، إذا ما أُحسن استخدامها عند حدود المعاني الجزئية وحدود المعاني الكلية. وهي علامات مهمة في تحليل النصوص حاسوبيًا.

٥. الخلاصة

لا تواجه جهود لسانيي الحاسوب العرب تحديات كبرى تعود إلى عدم كفاءة نظام كتابة العربية في تمثيل صوت الكلمة العربية أو صرفها. ومع ذلك قصرت جهودهم عن تطوير نظم التحليل اللغوي التي تضاهي النظم التي طورت للغة الإنجليزية. ولما تظهر حتى الآن تطبيقات الذكاء الاصطناعي العربية في الفهم الآلي التي يمكن توظيفها في تطبيقات الترجمة الآلية والبحث عن النصوص والتعرف على الحروف، وغيرها من التطبيقات. إن هذه النتيجة تدعو إلى دراسة أسباب هذا القصور ومعالجة أسبابه.

٦. المراجع

1. Abderahman, W. (1991) A concise dictionary of scientific roots: Graeco – Latin – English – Arabic, A new approach to the study of some aspects of Neologisation, Librairie du Liban.

٢. غانم قدوري الحمد، "علم الكتابة العربية"، دار عمار، عمّان، ٢٠٠٤، ص ١٠٥.

٣. محمد سالم العوفي: تطور كتابة المصحف الشريف وطباعته، مجمع الملك فهد لطباعة المصحف الشريف، المدينة المنورة . ١٤٢١هـ–٢٠٠٠م

٤. محمـد أحمد أبو عيد، مجلـة دراسات في اللّغة العربيـة وآدابها، فصلية محكّمة، العدد السادس عشر، شتاء / ٢٠١٤م.

٥. تمـام حسان، "اللغـة العربية معناهـا ومبناها"، عام الكتـب، القاهرة، ٢٠٠٤، ص ٤٠.

٦. ابـن جني، سـر صناعة الإعـراب، بتحقيق حسـن هنـداوي، دار القلم، دمشق،١٩٨٥ ج١ ص ١٧.

٧. كمال بشر، علم الأصوات، دار غريب، القاهرة، ٢٠٠٠م، ص ١٦٤.

أبحاث جامعة ليدز في مجال لسانيات المدونات العربية

د. إيريك أتويل (Eric Atwell) (*)

E.S.Atwell@leeds.ac.uk

أ. عبدالله بن يحيى الفيفي (**)

ayjfaifi@imamu.edu.sa

(*) أستاذ مشارك في قسم الحاسب الآلي في جامعة ليدز ببريطانيا.
متخصص في معالجة اللغة طبيعية، ومهتم بمجال التنقيب في النصوص وتحليل البيانات، وله اهتمام كذلك باللغة العربية، وقد أجرى العديد من الأبحاث حول هذه الموضوعات، كما أشرف على مجموعة من طلاب الدكتوراه الذين أسهموا في بحث هذه الموضوعات.

(**) معهد تعليم اللغة العربية، جامعة الإمام محمد بن سعود الإسلامية، الرياض.
متخصص في معالجة اللغة الطبيعية، شارك في عدد من المشاريع العلمية في مجال معالجة اللغة العربية حاسوبيًا، له عدة أبحاث منشورة حول مدونات المتعلمين، إضافة إلى مشاركته في تحكيم عدد من الأبحاث العلمية. أنشأ المدونة اللغوية لمتعلمي اللغة العربية (www.arabiclearnercorpus.com)، إضافة إلى بعض التطبيقات الحاسوبية في مجال المدونات اللغوية، مهتم بعلم اللغة الحاسوبي، وتعليم اللغة بمساعدة الحاسب.

الملخص

طـور الباحثون في مجال لسانيات المدونة ومعالجة اللغة الطبيعية في بريطانيا الكثـير من المدونات اللغويـة والأدوات الحاسوبية للبحـث في اللغة الإنجليزية. وفي جامعـة ليدز□□□، أردنا أن تمتد هذه الأبحـاث لتشمل اللغة العربية أيضًا؛ ولأن هذا يتطلـب بنـاء مدونات لغويـة عربية فقـد أنشأنا عـدة مدونات منهـا: مدونة تعليم العربية بواسطة الحاسب الآلي، والمدونة اللغوية للعربية المعاصرة، ومدونة الإنترنت العربيـة، ومدونـة اللغة العربية حول العالم، والمدونة اللغويـة لتدريس معلومات عن الإسـلام، والبنك الشجري للخطاب العربي، والمدونة اللغوية لمتعلمي اللغة العربية، والمدونة العربية لنصوص القـرآن الكريم، ومدونة الإحالة الثنائية لضمائر القرآن الكريم، ومدونـة الترابط الدلالي بين آيات القرآن الكـريم، ومدونة القرآن الكريم الموسومـة بموضوع الآيات مع الترجمـة الإنجليزية، ومدونة جامعة الملك سعود للغة العربيـة الفصحى. كذلك طورنا مجموعة مـن الأدوات لتحليل النصوص العربية، منهـا: الكشاف السياقـي aConCorde للغة العربية، والمحلـل الصرفي SALMA لوسم أقسام الكلام في المدونات العربية، وجدول وسم الأخطاء في مدونات المتعلمين العربيـة، والوسم الصوتـي والمقطعي للغة العربية، وتدريـب برامج التخاطب الآلي على مدونة عربية، والوسم الدلالي والتمثيل المعرفي للقرآن الكريم. وقد استُخدمت هـذه المدونـات اللغويـة والأدوات الحاسوبية لإجراء العديد مـن الأبحاث في مجال

□□□ جامعـة ليدز هي أقدم وأكبر جامعة في مقاطعـة Yorkshire في إنجلترا. افـتُتِحت الجامعة عام ١٩٠٤، ويعمـل فيها حاليًا ما يربو على سبعـة آلاف أكاديمي وموظف، كما يدرس فيها أكثر من اثنين وثلاثـين ألـف طالب من ١٤٦ دولة. تتميز الجامعة كذلك بأبحاثهـا الواسعة في عدد من الموضوعات الحيويـة كالهندسـة، والطب، واللغة، وغيرهـا. ويعد مركز "اللغة في جامعة ليـدز" واحدًا من أكبر المراكـز البريطانيـة التي تقدم أبحاثا ومنحا في الجوانب المتعلقة بدراسـة اللغة، وينتمي لهذا المركز أكـثر من مئة باحث في العلـوم الإنسانيـة، والآداب، والتربية، والعلوم الاجتماعيـة والنفسية، وعلوم الأحياء، والحاسب الآلي، والهندسة.

لسانيات المدونات العربية، ومن ذلك على سبيل المثال تعليم اللغة العربية، والمقارنة بين اللغتين الإنجليزية والعربية. ويعد موقع المدونة العربية لنصوص القرآن الكريم أحد أبرز الأمثلة على هذه الأبحاث، حيث يُستخدم على نطاق واسع من قبل الباحثين في اللغويات العربية، وكذلك علماء الشريعة الإسلامية، وعامة الناس في البلدان الإسلامية وغيرها من البلدان حول العالم. وقد قادتنا هذه الأبحاث إلى اقتراح موضوع "فهم النصوص الدينية" باعتباره أحد التحديات الجديدة والهامة في مجال البحث في لسانيات المدونات اللغوية.

١ - المقدمة

يمكن تعريف المدونة اللغوية (corpus) بأنها مجموعة من النصوص اللغوية -أو الأحاديث الشفهية- الطبيعية، التي تجمع لأغراض محددة، وتحفظ بطريقة قابلة للقراءة والبحث حاسوبيًا (McEnery et al., 2006; Leech, 1992; Jurafsky and Martin، ٢٠٠٨). وتتعدد التسميات العربية للفظة corpus، لكن تسميتها "مدونة لغوية" قد تكون الأكثر شيوعًا حسبما يرى محمود إسماعيل صالح (العصيمي، قيد النشر)، وكذلك وردت في مجموعة من معاجم ترجمة المصطلحات اللغوية والحاسوبية (انظر مثلاً البعلبكي، ١٩٩٠: ١٢٨، والمبارك، ١٩٩٥: ٦٧، والزهيري، ٢٠٠٦: ٣٦٥، والفهري، ٢٠٠٩: ٦٤).

لم تعد خافية على الباحثين اللغويين تلك الأهمية التي بات يحظى بها الجانب الحاسوبي في الدراسات اللغوية، فمثلًا يرى Kilgarriff (٢٠٠٧) أن أدوات معالجة اللغة الطبيعية وكذلك اللغويات الحاسوبية قد دخلت إلى مجالات البحث اللغوي كالمتنمر في باحة المدرسة، مما جعل كل ما هو غير حاسوبي إما خاضعًا لها، أو متواطئًا معها، أو مُهمّشًا. وقد أدرك الباحثون في جامعة ليدز هذه الأهمية، فطوّروا أدوات حاسوبية ومصادر لغوية بداية باللغة الإنجليزية، ثم توسعوا في هذه الأدوات والمصادر لتشمل عددًا من اللغات ومنها اللغة العربية. وفيما يتعلق بلسانيات

المدونـات العربيـة فثمة عـدة وحدات بحثيـة في الجامعة تشترك جميعًـا في العمل علـى هذا المجال، ومنها قسم الحاسب الآلي، وقسم اللغات والثقافات والمجتمعات، وقسم الدراسات العربية والإسلاميـة والشرق أوسطية، ومركز دراسات الترجمة، وقسـم اللغويات والصوتيات، ومعهد الذكاء الاصطناعي والأنظمة الحيوية. وتتبنى هـذ الأقسـام في جامعة ليـدز عددًا من المشاريـع البحثية في مجـالات رئيسة مثل: بنـاء المدونات العربية، وتطويـر أدوات تحليل النصوص العربية، إضافة إلى إجراء الدراسات العلمية في مجال لسانيات المدونات العربية.

نسـرد في الفصـل الثاني من هذا البحـث ثبتًا بأهم المدونـات اللغوية العربية التـي أُنشئت في جامعة ليدز، ثـم نخصص الفصل الثالث لاستعراض مجموعة من الأدوات الحاسوبيـة المصممة لتحليل النصوص العربية، وفي الفصل الرابع نتحدث عـن أبرز الموضوعـات البحثية التي استفادت مـن هذه المدونـات اللغوية والأدوات الحاسوبية، ثم نختم بأهم مجالات البحث اللغوي التي نسعى لدراستها مستقبلًا.

٢- المدونات اللغوية العربية

أنشئـت في جامعة ليدز مجموعة من المدونات اللغويـة العربية لتكون مصدرًا مساعـدًا لإجراء الدراسات علـى اللغة العربية، ونسرد فيما يلـي هذه المدونات مع إيراد أمثلة للأبحاث التي استفادت من كل واحدة منها.

مدونة تعليم العربية بواسطة الحاسب الآلي (ABC: Arabic By Computer)

بـدأت مشاركـة جامعة ليدز في أبحـاث لسانيات المدونـات العربية من خلال مشـروع مدونة النصوص العربية الفصحى لتعليم اللغة العربية بمساعدة الحاسب Arabic By Computer. وكان الهدف من هذا المشروع إنشاء مصدر لغوي لمتعلمي اللغـة العربية، يحتـوي على قاعـدة بيانات للنصوص العربية ومعجـم للمفردات (Brockett et al ١٩٨٩). وعنـد إنشـاء المدونـة، كان عرض النصوص العربيـة وتحريرهـا يتطلـب أجهزة وبرامج خاصة للحاسب الآلي مـن شركة أبل ماكنتوش،

وفي ثمانينات القرن الماضي لم تكن إدارة الحاسب الآلي في جامعة ليدز توفر حاسبات لتعليم اللغة، حيث كان استخدام هذه الأجهزة مقصورًا على أبحاث العلوم والهندسة. ومن هنا رأينا أهمية إيجاد وسائل تمكن المختصين من الوصول إلى المدونات العربية بسهولة ودون أية قيود لكونها مصدرًا لغويًا مهمًا في البحث والتعليم.

المدونة اللغوية للعربية المعاصرة (CCA: the Corpus of Contemporary Arabic)

كانت أبحاث لسانيات المدونات في جامعة ليدز قد طَوَّرتْ في بداياتها بعض المدونات وأدوات التحليل الخاصة باللغة الإنجليزية، ولكي يتم التوسع في هذه الأدوات لتشمل اللغة العربية فقد قامت بإنشاء أول مدونة لغوية عربية وإتاحة تنزيلها مجانًا، وتشمل مليون كلمة من اللغة العربية المعاصرة (Al-Sulaitiand Atwell ٢٠٠٥، ٢٠٠٦). صممت هذه المدونة لتكون قابلة للمقارنة مع مدونة Lan-caster-Oslo-Bergen Corpus (LOB) للإنجليزية البريطانية المعاصرة التي تشتمل على مليون كلمة، وكذلك مدونة Brown للإنجليزية الأمريكية المعاصرة، وفيها مليون كلمة أيضًا (Leech et al1983a). وعوضًا عن استنساخ نفس الأنواع الأدبية للنصوص الموجودة في هاتين المدونتين، فقد استُطلعت آراء مجموعة من المستخدمين المحتملين لهذه المدونة في مجالي تعليم اللغة وتحليل النصوص العربية، للتعرف على مدى تفضيلهم لأنواع أدبية محددة، ومن ثم ضُمنت هذه الأنواع في المدونة. وقد استُخدمت المدونة اللغوية للعربية المعاصرة من قبل عدد من الباحثين في مجال لسانيات المدونات العربية وذلك لعدة أغراض بحثية، منها على سبيل المثال: تعليم الهجاء والمفردات في اللغة العربية (Erradi et al ٢٠١٢)، وتصنيف المعاجم العربية (Attia et al ٢٠١١)، وعمل معجم لما يكتب متصلا بالعربية (Elarian and Idris، ٢٠١٠)، وترجمة المجازات ذات الصبغة الثقافية في المقالات العلمية (Merakchi and Rogers ٢٠١٣)، والاختلافات المعجمية في قسمي الشؤون الدولية والرياضة في الصحف العربية (Abdul Razak ٢٠١١)،

ومنهـا أيضًـا الدراسات القائمة على المدونات في علم اللغـة الاجتماعي (Friginal and Hardy ٢٠١٤).

مدونة الإنترنت العربية (Arabic Internet Corpus)

كانـت المدونـة الوطنيـة البريطانيـة (BNC: British National Corpus) في تسعينـات القـرن الماضـي تمثل معيـارًا مُعتَمَـدًا في مجـال لسانيـات المدونات الإنجليزيـة، ولم يكـن هناك تمويل وخبرات كافية لبنـاء مدونات عامة وكبيرة لأي مـن اللغات الأخرى لتحاكي المدونة الوطنيـة البريطانية مئة مليون كلمة). وللتغلب على ذلك فقد طُوِّرتْ منهجية لاستخدام محتويات شبكة الإنترنت باعتبارها مدونة لغويـة (Baroni and Bernardini ٢٠٠٤)، ممـا طلـب تحديد قائمـة من الكلمات تمثـل اللغة الهدف وتُستخدَم من قبل برامج خاصة تقوم بجمع مواد المدونة، وذلك بإرسـال مجموعة من هذه الكلمات للبحث عنها في محـركات البحث المعروفة على شبكـة الإنترنت، مثـل Google وYahoo وBing وغيرها، ومـن ثم تنزيل صفحات النتائـج، وتصفية النصوص، ثم تجميعها في مدونة واحدة. وفي جامعة ليدز اتُّبِعَتْ هذه المنهجيـة (باعتبار الإنترنت مدونة لغوية كبرى) لجمـع مدونات من الإنترنت للغة العربية، والصينية، والفرنسية، والألمانية، والإيطالية، والإسبانية، والبولندية، والروسية (Sharoff ٢٠٠٦). وهي مدونات متاحة للجميع من خلال واجهة لكشاف السياقـات والمتصاحبات على الموقـع التـالي: http://corpus.leeds.ac.uk/inter-net.html. وتشمـل مدونة الإنترنت العربيـة ١٧٦ مليون كلمة، وقد أضيفت الأصول المعجميـة إلى كلماتها في وقت لاحق باستخدام برنامج التحليل الصرفي SALMA (Sawalha and Atwell ٢٠١٣a).

مدونة اللغة العربية حول العالم (World Wide Arabic Corpus)

جمـع الباحثون في جامعة ليدز مدونـة لغوية عربية من أنحاء العالم (Atwell et al ٢٠٠٧) مماثلـة لمدونـة اللغة الإنجليزية حول العـالم World Wide English

Corpus. وتشمـل مدونات فرعية تتألف كل واحدة من مئتي ألف كلمة من كل دولة، وذلـك لدراسـة اختـلاف اللهجات من بلـد إلى بلد. وقـد استُخدمت هـذه المدونة لدراسـة التنوع بـين لغات الأقطار العربيـة (أو ما يعرف باللهجـات) في استخدام أدوات الوصـل (Hassan et al ٢٠١٠، ٢٠١٣)، وكذلـك التبايـن في اللغـة العربية وإنجليزية العرب في العالم العربي (Atwell et al ٢٠٠٩).

المدونة اللغوية لتدريس معلومات عن الإسلام (Corpus for Teaching about Islam)

استُخدمت شبكة الإنترنت - باعتبارها مدونة لغوية - لجمع مدونة تخصصية تضم النصوص المستخدمة لتدريس معلومات حول الإسلام للمستوى الجامعي (At-well et al ٢٠١١)، وذلك من أجل تأليف موسوعة جامعية لاستخدامها في تدريس الطلاب عن الإسلام والمسلمين، وتشمل اللغة، واللسانيات، ودراسات المناطق.

البنك الشجري للخطاب العربي (Arabic Discourse Treebank)

يحتاج تحليل الخطاب في اللغة العربية إلى نوع مختلف من وسم الكلمات (وهو مـا يعرف بالتحشية annotation)، وقد أنشـأت الباحثتان Al-Saif and Markert (٢٠١٠) البنـك الشجري للخطـاب العربي، وهي مدونة تشمـل ٥٣٧ نصًا إخباريًا قـد وُسِمَتْ جميـع أدوات الوصل فيهـا، إضافة إلى العلاقة التي تربـط كل أداة مع طرفيها. وقد تطلب هذا تطوير أداة حاسوبية لوسم الخطاب في النصوص العربية، وكذلك إنشاء موقع على شبكة الإنترنت لنشر المدونة.

المدونة اللغوية لمتعلمي اللغة العربية (ALC: Arabic Learner Corpus)

أُنشِئَـت المدونة اللغوية لمتعلمي اللغة العربية لتكون مصدرًا لغويًا لأبحاث تعليم اللغـة العربية، وكذلك للبحث في مجال معالجة اللغـة الطبيعية (Alfaifi and At-well 2013a, Alfaifi et al 2014). يمكن تنزيل نصوص المدونة من شبكة الإنترنت مجانًـا، وهي تتألف مـن مجموعة من المواد المكتوبة والمنطوقـة التي حررها متعلمو اللغة العربية في المملكة العربية السعودية خلال العامين ٢٠١٢ و٢٠١٣، وتضم ١٥٨٥

نصًـا (٢٨٢،٧٣٢ كلمـة)، شارك في تحريرها ٩٤٢ طالبًا مـن ٦٧ جنسية، و٦٦ لغة أم مختلفـة، ويبلغ متوسط طـول النص الواحد ١٧٨ كلمة. وتضـم نصوص المدونة نوعـين أدبيين: الأول سرديّ يحكي فيه الطالب رحلتـه في إحدى الإجازات، والثاني يناقش فيه الطالب اهتماماتـه الدراسية ومستقبله العلمي. ويستطيع الباحثون من خلال المعلومات الوصفية للطلاب والنصوص (metadata) التعرف على خصائص النص اللغـوي وكذلك مؤلف النص، مما يضيف عمقًا أكثر لتحليل بيانات المدونة، كمـا تفيد أيضًـا في مجالي التعرف الآلي على النص المكتـوب بخط اليد، والتعرف علـى الـكلام المنطوق. كمـا أن أصـول الأوراق - المكتوبة يدويًا مـن قبل الطلاب - متاحـة للتنزيـل بعـد أن أُدخِلتْ عن طريـق الماسح الضوئي، ثـم حفظت في ملفات بصيغة PDF. وكذلك الحال بالنسبـة للتسجيلات الصوتية لأولئك الطلاب الذين أذنوا بنشرها على الإنترنت، فهي متاحة للتنزيل في صيغة صوتية رقمية MP3 (ما مجموعه ثلاث ساعات ونصف تقريبًا).

سُمِّيَـتْ ملفات المدونـة بطريقة تسمـح بالتعرف على الخصائـص الأساسية للنـص وكذلك المؤلف، نحو: S038_T2_M_Pre_NNAS_W_C، وهي بالترتيب من اليسار: رقم الطالب، رقم النص، جنس الطالب، المرحلة العامة، ناطق بالعربية باعتبارها لغته الأم أو ناطق بغيرها، نوع النص (مكتوب أم منطوق)، مكان تحرير أو تسجيل النص (في الصف أو في المنزل). يمكن تنزيل نصوص المدونة من موقعها الرئيـس (http://www.arabiclearnercorpus.com)، كما يمكن البحث فيها عن طريـق الإنترنت من خلال الموقع التـالي (http://www.alcsearch.com) أو موقع SketchEngine (Kilgarriff et al ٢٠١٤). استُخدمت هذه المدونة في تقنية تهدف للتعرف على النصوص التي حررها الناطقون باللغة العربية وتمييزها عن تلك التي حررهـا الناطقون بغـير العربيـة (Malmasi and Dras ٢٠١٤)، وكذلك تصحيح الأخطاء آليًا (Mohit et al ٢٠١٤).

المدونة العربية للغة القرآن الكريم (Quranic Arabic Corpus)

بعـد إصدار المدونـة اللغوية للعربية المعاصـرة (Corpus of Contemporary Arabic) ظهـرت عدة أنواع من المدونات العربية المتاحة للاستخدام، لكن نصوص القـرآن الكـريم، وكذلك نصوص اللغة العربيـة الفصحى، لم تجد نفس العناية مـن قبـل المهتمين ببنـاء المدونات العربيـة. ولذا تعد المدونـة العربية للغـة القرآن الكـريم Quranic Arabic Corpus (Dukes et al 2013) مـن أشهـر المشاريع في هـذا الجانب، فقـد بُنـيَ هذا المصدر اللغوي من خـلال شراكة انطلقت من جامعة ليـدز. وتحوي هذه المدونـة عدة طبقات من الوسم، مثل أقسـام الكلام بعد تجزئة الكلمات بنـاء على الوحدات الصرفيـة (Dukes and Habash, 2010)، والتحليل النحـوي القائـم علـى التوابـع (Dukes and Buckwalter, 2010, Dukes et al 2010)، والمعـاني الإنجليزية لمفردات القرآن العربية، وكذلك الترجمة الإنجليزية للآيـات، وتضم المدونة أيضًا تسجيلات صوتيـة لتلاوة الآيات، إضافة إلى تصنيف الموضوعـات في القرآن الكريم. وكان الدافع لهذا المشروع إنتاج مصدر يساعد على فهم القـرآن الكريم، وإجراء المزيد من الأبحاث علـى نصوصه. ويمكن القول بأنه يختلف عن المشاريع الأخرى بتوفيره مادة لغوية أكثر عمقًا، تقوم على تحليل قواعد اللغـة العربية حسب المنهج العربي المعروف بالإعراب. وقد بات بالإمكان من خلال تبنـي هذا النهج تحفيز الباحثين اللغويين والشرعيـين للعمل على موضوع التحشية بأسلـوب تعاوني عن طريق الإنترنت، حيث يُستخدم الوسم الآلي المبني على قواعد محـددة في هـذه الطريقة الجديدة لتحشيـة المدونة لغويًا، مع إجـراء تدقيق يدوي أولي، ثـم القيـام بالتصحيح عن طريق الإنترنت من قبـل الباحثين المتعاونين. وقد استفـاد الوسـم الصرفي لمدونـة العربية لنصوص القرآن الكـريم من الجهد الذي قدمه مئة متطوع تقريبًا، وذلك على شكل اقتراحات لتدقيق الوسم اللغوي للمدونة. كما كان هناك دور إشرافي لعدد قليل من المختصين في عملية الوسم، بحيث يقبلون اقتراحـات التصحيح المقدمـة عن طريق المتعاونـين أو يرفضونها. كمـا استفادت

المدونـة كذلك من الكم الكبير للتراث النحوي العربي الذي قُدّم من خلال تعليقات الباحثين على نصوص المدونة.

التحدي الأبرز الذي واجـه عملية تحشية نصوص القرآن عن طريق الإنترنت تمثل في الحاجة إلى بناء برنامج خاص، عبارة عن منصة عمل تساعد على الوسم بأسلـوب تعاوني، ومن هنا بدأت فكرة بناء منصـة التحليل اللغوي متعدد الوسائط (LAMP: Linguistic Analysis Multimodal Platform) (Dukes and Atwell ٢٠١٢). استخدمـت المدونة العربية لنصوص القرآن الكريم كمصدر معياري لعـدد من الأبحـاث التي أجريت على اللغـة العربية الفصحى، ومـن هذه الأبحاث استخـراج جذور الكلمـات العربيـة (Yusof et al ٢٠١٠)، والتحليـل النحوي للغة العربيـة (Mohammed and Omar 2011, Rabiee 2011)، واستخـراج البصمة الأسلوبيـة في اللغـة العربيـة (Alqurneh et al forthcoming)، وتحليـل الترابط في دراسـات الترجمة العربيـة (Tabrizi and Mahmud ٢٠١٣)، والتلخيص الآلي (El-Haj et al forthcoming)، وتحليل الصيغ الشفهية (Bannister ٢٠١٤). كما كان للمدونـة أثر اجتماعي كبير تمثل في مليون زيارة لموقع المدونة خلال عام واحد، ومن ضمن هؤلاء الـزوار ناطقون بلغات غير العربية قادتهم للاستفادة من المدونة رغبتهم في الوصـول إلى فهم أعمق لنص أصيل من اللغة العربية الفصحى، وذلك بالاستفادة من التحشية اللغوية المستخدمة في المدونة.

مدونة الإحالة الثنائية لضـمائر القرآن الكريم (QurAna: Quran Pronoun Anaphoric Co-Reference Corpus)

تعـد مدونة الإحالـة الثنائية لضمائـر القرآن الكـريم (Sharaf and Atwell 2012a) مدونـة غنية من حيـث التحشية، فقد وسمت جميـع الضمائر الشخصية فيها بمعلومات عن العائد الذي يشير إليه الضمير، سواء أكان العائد سابقًا للضمير أم كان لاحقًـا لـه وهو الغالـب في المدونة. كما شمل الوسم كذلـك العائد نفسه، أي

الكلمــة أو الجملة التــي تحوي المعنى، أو الشخص، أو الكيــان، أو الفكرة التي يشير إليهــا الضمير، ولهــذا سميت بمدونة الإحالــة الثنائية لشمــول الضمير وعائده في التحشيــة. في هــذه المدونة وَسَم الباحثــون ما يزيد عــن ٢٤.٥٠٠ ضمير بمعلومات حــول العائد، وكذلك أكثر من ألف عائد - تتكون مــن أسماء وعبارات - بمعلومات عن الأشخاص، أو الكيانات، أو المفاهيم، وذلك لجميع الأسماء أو الجمل التي تشير إليهــا هذه الضمائر. لم يكن من السهل تحديد العائــد لجميع الضمائر الشخصية في القرآن الكــريم، لكن كتب التفسير وكذلك كلام العلماء حول هذه الضمائر كانا بمثابــة الدليل للقائمين على تحشية المدونة. وفي بعض الحالات الخاصة حيث يكون للضمــير أكــثر مــن عائد، فقد اتبعنــا ما ورد في تفســير ابن كثير الذي يعــد مؤلفًا مرجعيًــا في تفسير القرآن الكريم، وقد اعتُمد عليه بدرجة كبيرة في تحديد العائد، مع الأخذ كذلك بدرجة الاتفاق بين العاملين على تحشية المدونة حول عائد الضمير. وتعــد هذه المدونة أول مدونــة للغة العربية الفصحى يمكن تنزيلها مجانًا، مع كونها موسومة بهذا النوع من المعلومات عن الضمير وعائده.

مدونــة الترابــط الدلالي بــين آيات القــرآن الكــريم (QurSim: Quran Verse Similarity Corpus)

مدونــة الترابــط الدلالي بــين آيــات القــرآن الكــريم (Sharaf and Atwell 2012b) تمثــل إضافة جديدة في مجــال التحشية اللغوية على النص القرآني، ففي هذه المدونة - المتاحة مجانًا للباحثين اللغوين المهتمين ببحث موضوع التشابه اللغوي والعلاقــات الدلالية بين آيات القرآن الكــريم - وسمَتْ الآيات التي بينها علاقة من ناحية الدلالة، وقد اعتمد الباحثون على تفسير ابن كثير في استخراج العلاقات بين آيات القرآن الكريم، حيث يشير ابن كثير إلى الآيات ذات الصلة في معرض تفسيره لــكل آيــة، وقد مكن ذلــك من استخراج العلاقــات المرجعية بين هــذه الآيات والتي زاد عددهــا عن ٧.٦٠٠ علاقة مشتركة. نشرت "مدونــة الترابط الدلالي بين آيات القــرآن الكريم" على الموقع التالي على شبكة الإنترنت (TextMiningTheQuran.

com) حيث يستطيع المستخدم رؤية شبكة العلاقات المباشرة وغير المباشرة لأي آية من آيات القرآن الكريم. وقد أظهر تحليل المدونة أن ٣٣٪ فقط من الآيات المرتبطة ببعضهـا تشـترك كلماتها في نفس الجذر، مما يدل علـى أن الارتباط بينها يتجاوز التطابـق المعجمـي للألفـاظ إلى المعنى الـدلالي والمجال المعرفي. يمكـن استخدام هـذه المدونة لاستخراج الموضوعات الواردة في القرآن الكريم وتصويرها من خلال العلاقـات الدلاليـة بين الآيـات (Panju 2014)، وبمـا أن ابن كثـير قد تحدث في تفسيره عن علاقات النص القرآني وهو نص عربي فصيح، فهذه العلاقات مصدر جيـد يمكن الاستفادة منه في أبحاث ترجمة معـاني القرآن الكريم، ومن ذلك على سبيل المثال أن الآيات التي بينها ارتباط في اللغة العربية ينبغي أن تبقى كذلك عند ترجمـة معـاني الآيات إلى أي لغة أخـرى، وبالتالي فقد تكون هـذه المدونة مصدرًا للبحث في ترجمة معاني القرآن الكريم للغات الأخرى، وخصوصًا في موضوع الآيات المرتبطة دلاليًا والعلاقة بينها.

مدونـة القـرآن الكـريم الموسـومة بموضـوع الآيـات مـع الترجمـة الإنجليزيـة (Qurany: Quran Corpus Annotated with English Translations and Verse Topics)

مدونـة القـرآن الكـريم الموسومة بموضـوع الآيات مـع الترجمـة الإنجليزية Qurany (Abbas 2009, Abbas and Atwell 2013) عبـارة عـن أداة ثنائيـة اللغـة (إنجليزي - عربي) ذات قدرة على تحسـين الدقة والاستجابة (recall and precision) عنـد البحث في موضوعات القـرآن الكريم، وقد أمكن تحقيق هذا من خـلال الجمـع بين عدة أنواع مـن التحشية، حيـث وسمت كل آية من آيـات القرآن الكريم بمعلومات عن الموضوع الدلالي الذي تحتويه، والذي استُخرج من "مصحف التجويـد"، وهـو مرجـع موثـوق يحـوي فهرسًا لمـا يقارب ألفًـا ومئة مـن المفاهيم والموضوعـات مـع الآيـات المرتبطة بها. وقد حـوَّل الباحثان فهرسـ هذا المصحف إلى تصنيـف يعرضهـا بصـورة شجري يمكـن استخدامه من خـلال موقع المدونة

على شبكة الإنترنت، حيث يستطيع المستخدم التنقل بين شجرة التصنيف لإيجاد الفكرة المطلوبة، ومن ثم الوصول إلى قائمة بالآيات المرتبطة بها. كما وسمَتْ كل آية بثماني ترجمات إنجليزية مشهورة لمعاني القرآن الكريم، فبات بالإمكان إيجاد الآيات القرآنية من خلال البحث في هذه الترجمات الإنجليزية باستخدام أي كلمة أو كلمات مفتاحية، كما يمكن للمستخدم مشاهدة مجموعة من مرادفات الكلمات المفتاحية المستخدمة في البحث، والاستفادة منها في توسيع نطاقه وبالتالي تحسين جودة النتائج. بإمكان المستخدم أيضًا الوصول إلى بيانات هذه المدونة وتنزيل محتواها على شكل صفحات مكتوبة بلغة النصوص المتشعبة HTML، وتشتمل كل صفحة من هذه الصفحات على آية واحدة مع ترجماتها الإنجليزية، وكذلك قائمة مصحف التجويد للموضوعات المرتبطة بها. تتوافق لغة HTML هذه مع معايير محرك البحث Google، حيث يمكنك قصر البحث في هذا المحرك على موقع إنترنت محدد باستخدام العبارة (site) مع كلمات البحث، فعلى سبيل المثال تقوم عبارة البحث التالية:

"prayer site:http://www.comp.leeds.ac.uk/nora/html" بإيجاد جميع الآيات التي ورد ذكر الصلاة في ترجمات معانيها، أو ورد ذلك في قائمة موضوعات مصحف التجويد.

مدونة جامعة الملك سعود للغة العربية الفصحى (KSUCCA King Saud University Corpus of Classical Arabic)

يمكن دراسة الأنماط المعجمية في القرآن الكريم باستخدام أي كشاف للسياقات متوافق مع اللغة العربية مثل aConCorde، لكن الباحثين اللغويين وصناع المعاجم يحتاجون إلى مدونات كبيرة لدراسة المتصاحبات اللغوية والنماذج السياقية لمفردات اللغة. وفي اللغة الإنجليزية على سبيل المثال، أنشئَتْ المدونة الوطنية البريطانية لتكون أساسًا للبحث المعجمي، وهي تحتوي على مئة مليون

كلمة، أما في اللغة العربية الفصحى فالنص القرآني يحتوي على خمسين ألف كلمة تقريبًا (بناء على طريقة احتساب حدود الكلمة)، ولأن البحث في الدلالة التوزيعية لمفردات المعجم يحتاج إلى عدد كبير نسبيًا من الأمثلة لكل كلمة أو متصاحب لغوي للتمكن من دراسته، ولأن كثيرًا من الكلمات والعبارات في القرآن الكريم لم ترد إلا في مواضع قليلة؛ فقد تعاونت جامعة ليدز مع جامعة الملك سعود لإنشاء مدونة تحوي خمسين مليون كلمة من اللغة العربية الفصحى القريبة من فترة نزول القرآن الكريم، وهذه المدونة هي مدونة جامعة الملك سعود للغة العربية الفصحى (Alrabiah et al 2013, 2014a,b). وقد سمحت لنا هذه المدونة بإيجاد كثير من الأمثلة والسياقات لكلمات وردت في القرآن الكريم، وهي مدونة يمكن تنزيل محتوياتها من موقعها الخاص على شبكة الإنترنت (http://ksucorpus.ksu.edu.sa)، كما يمكن البحث فيها باستخدام موقع SketchEngine (Kilgarriff et al 2014). هذه المدونة تعد مفتاحًا للدراسات اللغوية التاريخية القائمة على المدونات (Alrabiah et al 2014a) وكذلك دراسات الدلالة التوزيعية لمفردات القرآن الكريم (Alrabiah et al 2014b).

٣. أدوات تحليل النصوص العربية

من المهم في ظل تنامي استخدام المدونات اللغوية وجود أدوات حاسوبية تستطيع الإفادة من هذه المصادر اللغوية، وبسبب الخصائص الفريدة للغة العربية فإنها تحتاج إلى أدوات تستطيع التعرف عليها ومعالجتها بطريقة صحيحة، وقد صمم الباحثون في جامعة ليدز مجموعة من الأدوات المخصصة لمعالجة اللغة العربية نستعرضها في هذا القسم.

الكشاف السياقي aConCorde **للغة العربية** (aConCorde Concordancer for Arabic)

من الملاحظ أن كشافات السياقات المتوفرة قبل عدة سنوات لم تكن مصممة

للتعامـل مـع الخصائـص المميزة للنصـ العربي، ومـن ذلك مثـلاً اختلاف رسم الحروف العربية عن تلك اللاتينية، والتباين في معايير تمثيلها حاسوبيًا، واختلاف رسم بعض الحروف بناء على موضعها في الكلمة، ومن تلك الخصائص أيضًا عدم رسـم الصوائـت القصيرة (الحـركات) في كثير من الأحيـان، والتباين في طريقة استخـدام علامات الترقيم، وكذلـك كتابة النص العربي من اليمـين إلى اليسار، والـذي أبرز الحاجة إلى تبديـل نافذتيّ "الكلمات السابقـة" و"الكلمات اللاحقة" في برامـج كشافات السياقـات لتلائم اتجاه النص العربي. ولهـذا فقد بُنِيَ كشاف السياقـات aConCorde في قسـم الحاسـب الآلي بجامعة ليدز، وهـو أداة مجانية ومفتوحـة المصدر، صممت في المقام الأول لمساعـدة الباحثين في لسانيات المدونات العربيـة (Roberts et al 2005, 2006). وقـد أشادت دراسـة علمية أجريت حول كشافـات السياقـات في ذلـك الوقـت (Wiechmann and Fuhs 2006) ببـرنامج aConcorde الـذي يوفر دعمًا كاملاً للنصوص العربية، ويتمثل هذا الدعم في عدة جوانـب منها واجهة المستخدم التي يمكن تحويلها بين اللغتين العربية والإنجليزية، وتبنيه لترميز الحروف المعياري Unicode، إضافة إلى دعمه لمجموعة من الخطوط العربيـة، وكذلك اتجاه النص العربي مـن اليمين إلى اليسار. وبعد عقد من الزمن أُجـريَ تقييـم آخر لـلأدوات المستخدمة في تحليل المدونات العربيـة (Alfaifi and Atwell 2014a) حيـث وجـدت الدراسـة أن معظـم الأدوات مـا زالـت تعاني من مشاكل في دعم اللغة العربية. وقد استُخدم aConcorde في عدة دراسات في مجال لسانيـات المدونـات العربية، ومن ذلـك الكشف عن الأنماط المعجميـة الشائعة في النصوص العربيـة (El-Haj et al forthcoming)، والكلمات والعبارات المفتاحية (Ali 2012)، والتعـرف علـى أساليب الجرائم من خلال مدونـة إخبارية لتقارير الجرائم (Alruily 2012).

المحلـل الصـرفي SALMA لوسـم أقسـام الـكلام في المدونـات العربيـة

(SALMA: Standard Arabic Language Morphological Analysis)

أسهـم الباحثـون في جامعـة ليدز في مشـروع وسم مدونـة LOB للإنجليزية البريطانيـة المعاصـرة (Atwell 1982, Leech et al 1983b)، وقـد منحهم ذلك خـبرة استفـادوا منها في وسـم المدونـات العربيـة (Atwell 2008، Atwell et al 2008)، حيـث جرى العمل على تطوير أداة للتحليـل الصرفي، وكذلك وسم أقسام الـكلام في المدونـات العربية، إضافـة إلى بعض الوظائف الأخـرى. وكانت البداية بوسـم عينة قصـيرة من القرآن الكـريم إضافـة إلى نص إخبـاري، لتكون هاتان العينتـان بمثابة معيار يمكن الاعتماد عليه لإجـراء مقارنة بين المحللات الصرفية الحاليـة (Sawalha and Atwell 2008). كمـا جرت مقارنـة بين منهجين لتحليل اللغـة العربيـة صرفيًا، يقوم الأول على أساس القواعـد اللغوية، بينما يعتمد الثاني علـى المدونات اللغويـة (Sawalha and Atwell 2009). قادت هذه الدراسات إلى تطويـر أداة جديدة أطلق عليهـا SALMA، وهي أداة لتحليـل الوحدات الصرفية الصغرى في اللغة العربية Arabic Corpus Part-of-Speech Tagging and Morphological Analysis، مع وسمها باستخدام جدول الوسم الصرفي لأقسام الكلام المضمن في هذه الأداة.

والاسـم SALMA يشير إلى: محلل ليدز الصرفي لصوالحة وأتويل Sawalha Atwell Leeds Morphological Analysis (Sawalha and Atwell 2013a) أو المحلـل الصرفي للغة العربية الفصحـى Standard Arabic Language Morphological Analysis (Sawalha et al 2013). ويشتمـل علـى مجموعـة من الأدوات لتحليـل المدونـات العربية، ومنها جدول معيـاري لوسم أقسام الـكلام، يأخذ بعين الاعتبار السمات الصرفية الأصيلة للغة العربية (Sawalha and Atwell 2013b)، وقد تضمن هذا العمل تحليلًا لأبحاث اللغوين العرب القدماء واستخدامها كأساس للجدول الحالي الذي يعطي تصنيفًا شاملًا ومفصلًا لبنية الكلمة في اللغة العربية.

ومع وجود عدد من الجداول الأخرى لوسم أقسام الكلام، والتي أُعدَّتْ للقيام بمهام محددة في اللغة العربية، إلا إنها في الغالب إما مستفادة من نماذج أعدت للغة الإنجليزية، أو إنها تغطي جزءًا محدودًا فقط من الدراسات الصرفية العربية، ولهذا فقد أنشئ مقياس لمقارنة وتقييم جداول أقسام الكلام كجزء من المشروع الحالي. وقد أضيف إلى أداة التحليل الصرفي وجدول أقسام الكلام معجم عربي كبير مستمد من المصادر المعجمية مفتوحة المصدر على شبكة الإنترنت، وكذلك المعاجم التراثية للغة العربية (Sawalha and Atwell 2010b)، كما أضيفت كذلك ميزة لعرض للتحليل الصرفي العربي بطريقة مرئية، حيث تظهر الوحدات الصرفية مميزة بألوان مختلفة عن بعضها ليسهل تمييزها بصريًا (Sawalha and Atwell 2012). وللتحقق من جودة عمل هذه الأدوات على المدونات الكبيرة، فقد تمت تجربتها على مدونة الإنترنت العربية التي يصل حجمها إلى ١٧٦ مليون كلمة (Sawalha and Atwell 2013a)، كما استُخدمتْ في مجموعة من الأبحاث في مجال لسانيات المدونات العربية، ومنها إنشاء قوائم بالمفردات اللغوية لتعليم اللغة العربية (Kilgarriff et al ٢٠١٤)، وتعليم الهجاء والمفردات العربية (Er-radi et al 2012)، وكذلك التحليل النحوي العربي (Rabiee 2011)، وأيضًا تحليل وسائل التواصل الاجتماعي العربية (El-Beltagy and Ali 2013).

جدول وسم الأخطاء في مدونات المتعلمين العربية (Arabic Learner Corpus Error Tag-Set)

أنشئ جدول تصنيف الأخطاء لاستخدامه في وسم الأخطاء اللغوية في مدونات المتعلمين العربية (Alfaifi et al 2013، Alfaifi and Atwell 2012، 2014b)، وقد استفاد هذا التصنيف من مجموعة من تصنيفات الأخطاء المصممة لمدونات المتعلمين، ليكون مناسبًا للأخطاء اللغوية التي يقع فيها متعلمو اللغة العربية. استخدُم هذا التصنيف لتطوير أداة هي عبارة عن محرر لوسم أخطاء الطلاب في المدونة اللغوية لمتعلمي اللغة العربية (Alfaifi and Atwell 2015)، والذي يمكن الاستفادة منه في وسم أي مدونة من مدونات المتعلمين.

الوسـم الصوتي والمقطعي للغة العربية (Arabic Phonetic and Prosodic Tagging)

نـوع آخر مـن أنوع الوسم يتمثـل في كتابة النصوص صوتيًـا – ومنها نصوص القـرآن الكـريم – وذلك لعـدة أهـداف وتطبيقـات كالتمكن من قراءتهـا بصوت مسموع، والتعـرف الآلي على الكلام المنطوق. يستخدم علمـاء الأصوات الألفبائية الصوتيـة الدوليـة IPA (International Phonetic Alphabet) لكتابـة النصوص المنطوقـة في مختلف لغات العالم، ولأن رسم اللغة العربية لا يعبر عن طريقة اللفظ بدقة كبيرة، إضافة إلى أنه لا يوجد ربط مباشر بين كل واحد من الحروف المكتوبة في اللغـة العربيـة مع رموز الألفبائيـة الصوتية الدولية API، فقـد بُنِيَ نظام دقيق للربـط بينهمـا مع كتابـة النصوص العربيـة صوتيًـا وبطريقة آلية، وقـد بني هذا النظام على تحليل لتلاوة القرآن الكريم وأحكام التجويد كالوقف والغنة ونحوهما، إضافة إلى الاستفادة من مجالي اللغويات العربية وعلم الأصوات الحديث (Brierley et al forthcoming, Sawalha et al 2014). ويتضمـن النصـ القـرآني على وجه الخصوص رموزًا مقطعية تشير إلى عدة أنواع من الوقف أو حدود الجملة التي ينبغـي مراعاتهـا عند قراءة الآيات بصوت مسموع، وهـي جزء من أحكام التجويد التي تضبـط عملية النطق والوقف أثناء تلاوة القـرآن الكريم. وقد استُخدم نظام الربـط بالألفبائية الصوتيـة الدولية، وكذلك رموز المقاطـع، في التحشية الصوتية والمقطعية للمدونة العربية للقرآن الكريم الموسومة بحدود الوقف في النص القرآني (Brierley et al 2012a،b). يمكـن أن تساعـد هـذه المدونة غـيرَ الناطقين باللغة العربيـة على تعلم التلاوة الصحيحـة لآيات القرآن، كما يمكن أن تُستخدم لتدريب أدوات الوسـم المقطعـي لنصوص اللغة العربية بما في ذلـك اللغة العربية المعاصرة (Sawalha et al 2012a,b).

تدريـب برامـج التخاطـب الآلي على مدونـة عربيـة (Arabic Corpus-Trained Chatbots)

مـن الاستخدامات المبتكرة للمدونـات استخدامها في تدريب برامج التخاطب الآلي عـن طريق الإنترنت، والتي تقـوم على أنظمة تعلم الآلة Machine Learning (Abu Shawar and Atwell 2005a)، ومن ثم تستخدم هذه البرامج بعد التدريب كأداة لاستكشـاف المدونات نفسها (Abu Shawar and Atwell 2005b). وكمثال علـى ذلك، فقد جـرى تدريب برنامـج للتخاطب الآلي على مدونـة قرآنية، وهو ما أنتج لنا نظام تخاطب عربي يعطي إجابات من القرآن الكريم (Abu Shawar and Atwell ٢٠٠٤)، كما دُرِّبتْ نسخة أخرى منه على مدونة للأسئلة الأكثر شيوعًا عن الحوسبة العربية، ليقوم بالإجابة على أي سؤال حول هذا الموضوع (Abu Shawar and Atwell 2009).

الوسـم الدلالي والتمثيل المعـرفي للغة القرآن الكـريم (Semantic Tagging and Knowledge Representation for Quranic Arabic).

تم إضافـة الكثـير من مستويـات الوسم اللغـوي للنص القرآني مثـل: أقسام الـكلام، والصـرف، والإحـالات المرجعية، والكتابـة الصوتية، والحـدود المقطعية للجمـل، والبنيـة النحوية، كما أضيفت عدة أنواع مـن التحشية التي تمثل معلومات معرفية في القرآن، ومنها تصنيف الكيانات الاسمية المعروفة بالضمائر الشخصية، وكذلك موضوعات الآيات، وعلاقاتها الدلالية، وأيضًا ترجمة معاني القرآن الكريم إلى اللغـة الإنجليزيـة، ومن هذه الترجمات ما هو علـى مستوى الكلمة ومنها ما هو على مستوى الآية. ويهدف الباحثون في جامعة ليدز إلى توحيد جميع هذه التحشيات اللغويـة والمعرفية (Abbas et al 2013, Abbas and Atwell 2013) لإنشاء تصور للتمثيل المعرفي في القرآن الكريم، وذلك للتمكن من إجراء وسم دلالي له (Sharaf and Atwell 2009، Alrehaili and Atwell 2013, 2014).

٤. البحث اللغوي باستخدام المدونات العربية وأدوات تحليل نصوصها

استُخدمت المدونات اللغوية والأدوات الحاسوبية التي طورها الباحثون في جامعة ليدز لإجراء العديد من الأبحاث في مجال لسانيات المدونات العربية، ومن ذلك تعليم اللغة العربية، والمقارنة بين اللغتين الإنجليزية والعربية، وهو ما سنتحدث عنه في هذا القسم.

تعليم اللغة العربية باستخدام المدونات

من خلال المشروع الرائد ABC (Arabic By Computer) بنى الباحثون قاعدة بيانات للنصوص العربية ومعجم للمفردات لاستخدامهما في تعليم وتعلم اللغة العربية (Brockett et al 1989). وفي الآونة الأخيرة استطعنا في جامعة ليدز تجربة كشاف للسياقات وبرامج للتخاطب الآلي، مستخدمين مدونات على شبكة الإنترنت، وذلك بهدف تعليم اللغة العربية (Al-Sulaiti et al 2005, 2007)، ومن خلال الأبحاث القائمة على المدونات أيضًا أنشئَتْ قوائم للمفردات اللغوية التي تخدم متعلمي اللغة العربية (Kilgarriff et al 2013)، وقد تواصلنا مع مجموعة من معلمي ومتعلمي اللغة العربية في قسم اللغة العربية بجامعة ليدز الذين يستخدمون المصادر اللغوية للمدونات العربية في تعلم وتعليم هذه اللغة. كما أن المجتمع المحلي للمسلمين أقام مدرسة لتعليم الأطفال اللغة العربية في أيام السبت من كل أسبوع، لتمكينهم من قراءة وفهم القرآن الكريم، وقد استمتع الطلاب بنظام التخاطب المبتكر على شبكة الإنترنت، والذي جرى تدريبه على مدونة قرآنية ليعطي إجابات بنفس اللغة العربية الفصحى المستخدمة في القرآن الكريم (Abu Shawar and Atwell 2004).

مقارنة اللغتين العربية والإنجليزية القائم على المدونات

كان البحث في جامعة ليدز قد بدأ بمجال لسانيات المدونات الإنجليزية، (انظر مثلًا Leech et al 1983a،b)، ثم تبعه اهتمام بإجراء مقارنات بين اللغتين

العربيــة والإنجليزية تقوم على مدونات لهاتين اللغتين، وتشمل هذه المقارنات تأثير اللغة العربية على إنجليزية العرب، والتنوع في استخدام اللغة الإنجليزية في العالم العربــي (Atwell et al 2009)، كما تشمل كذلك جـداول الوسم الصرفي وأقسام الـكلام في اللغتين العربية والإنجليزيـة (Atwell 2008, Sawalha and Atwell 2013c)، وأيضًــا التمثيـل البصــري للنــبر والوقـف في نصوص المدونــات العربية والإنجليزيــة من خلال تمييز مواضعها في النص بعلامات متباينة الألوان (Brier-ley et al 2012c).

الخاتمة

تستخـدم أجهزة الحاسب الآلي والمدونـات اللغوية على نطاق واسع في أبحاث اللسانيـات، وقد طوّر الباحثون في جامعة ليدز عدة مدونات لغوية وبرامج حاسوبية علـى شبكة الإنترنـت للاستفادة منها في مجـال لسانيات المدونـات العربية، وهي أدوات مفتوحـة المصـدر ومتاحـة علـى شبكة الإنترنـت وليست تجاريـة، آملين أن يسهـم ذلك في استخدامها علـى نطاق واسع. ويعتبر وسم وتمثيل الجوانب الدلالية والمعرفيـة في النصوص العربيـة – وخصوصًا في القرآن الكريم والنصوص الدينية الأخـرى – تحديًا للأبحاث القادمة، كما يمثل فهم القرآن حاسوبيًا التحدي الأكبر في مجال لسانيات المدونات العربية.

المراجع

١. البعلبكي، رمزي منير (١٩٩٠) معجم المصطلحات اللغوية (إنجليزي عربي). دار العلم للملايين، بيروت.

٢. الزهـيري، نبيـل (٢٠٠٦) معجم المصطلحات اللغويـات في المعلوماتية. مكتبة لبنان، بيروت.

٣. صالـح، محمود إسماعيل (قيد النشر) المدونـات اللغوية وكيفية الإفادة منها، تحرير صالح العصيمي، المدونات اللغوية العربية: بناؤها وطرائق الإفادة منها. مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية، الرياض.

٤. الفهـري، عبدالقادر الفاسي (٢٠٠٩) معجم المصطلحات اللسانية (إنجليزي - فرنسي- عربي). دار الكتاب الجديد، بيروت.

٥. المبـارك، مبارك (١٩٩٥) معجم المصطلحـات اللغوية (فرنسي - إنجليزي - عربي). دار الفكر، بيروت.

6. Abbas، N and E Atwell. 2013. ‘Annotating the Arabic Quran with a classical semantic ontology.’ Proceedings of WACL’2 Second Workshop on Arabic Corpus Linguistics.

7. Abbas، N، L Aldhubayi، H Al-Khalifa H، Z Alqassem، E Atwell، K Dukes، M Sawalha، and M Sharaf. 2013. ‘Unifying linguistic annotations and ontologies for the Arabic Quran.’ Proceedings of WACL2 Second Workshop on Arabic Corpus Linguistics.

8. Abbas، N. 2009. 'Quran Search for a Concept Tool and Website'. MRes Thesis، School of Computing، University of Leeds.

9. Abdul Razak، Z. 2011. 'Modern media Arabic: a study of word frequency in world affairs and sports sections in Arabic newspapers.' PhD Thesis، University of Birmingham.

10. Abu Shawar، B and E Atwell. 2004. 'An Arabic chatbot giving answers from the Quran.' Proc TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles.

11. Abu Shawar، B and E Atwell. 2005a. 'Using corpora in machine-learning chatbot systems.' International Journal of Corpus Linguistics، vol. 10، pp. 489-516.

12. Abu Shawar، B and E Atwell. 2005b. 'A chatbot system as a tool to animate a corpus.' ICAME Journal: International Computer Archive of Modern and Medieval English Journal، vol. 29، pp.5-24.

13. Abu Shawar، B and E Atwell. 2009. 'Arabic Question-Answering via Instance Based Learning from an FAQ Corpus.' Proceedings of CL2009 Corpus Linguistics.

14. Al-Saif، A، and K Markert. 2010. 'The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic.' Proceedings of LREC'2010: Language Resources and Evaluation Conference.

15. Al-Sulaiti، L and E Atwell. 2006. 'The design of a corpus of contemporary Arabic.' International Journal of Corpus Linguistics، vol. 11، pp. 135-171.

16. Al-Sulaiti، L، A Roberts، and E Atwell. 2005. 'The use of corpora and concordance in the teaching of contemporary Arabic.' Proceedings of EuroCALL'2005.

17. Al-Sulaiti، L، A Roberts، B Abu Shawar، and E Atwell. 2007. 'The Use of Corpus، Concordancer and Chatbot in the Teaching of Contemporary Arabic.' Proceedings of CL›2007 Corpus Linguistics

18. Al-Sulaiti، L، and E Atwell. 2005. 'Extending the Corpus of Contemporary Arabic.' Proceedings of CL›2005 Corpus Linguistics.

19. Alfaifi، A and E Atwell. 2012. 'Arabic Learner Corpora (ALC): A Taxonomy of Coding Errors'. Proceedings of ICCA'2012 International Computing Conference in Arabic.

20. Alfaifi، A and E Atwell. 2013a. 'Arabic Learner Corpus v1: A New Resource for Arabic Language Research.' Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics.

21. Alfaifi، A and E Atwell. 2013b. 'Arabic Learner Corpus: Texts Transcription and Files Format.' Proceedings of CORPORA'2013 International Conference on Corpus Linguistics.

22. Alfaifi، A and E Atwell. 2014a. 'Tools for Searching and Analysing Arabic Corpora: an Evaluation Study.' Proceedings BAAL-CUP'2014 British Association for Applied Linguistics and Cambridge University Press Applied Linguistics Workshop.

23. Alfaifi، A and E Atwell. 2014b. 'An evaluation of the Arabic error tagset v2.' Proceedings of AACL'2014 American Association for Corpus Linguistics.

24. Alfaifi، Abdullah and Atwell، Eric. 2015. Computer-Aided Error Annotation A New Tool for Annotating Arabic Error. The 8th Saudi Students Conference، 31 January – 1 February 2015، London، UK.

25. Alfaifi، A، E Atwell، and G Abuhakema. 2013. 'Error Annotation of the Arabic Learner Corpus: A New Error Tagset. Language Processing and Knowledge in the Web، vol. 8105، pp.14-22. Springer.

26. Alfaifi، A، E Atwell، and I Hedaya. 2014. 'Arabic Learner Corpus (ALC) v2: A New Written and Spoken Corpus of Arabic Learners.' Proceedings of LCSAW'2014 Learner Corpus Studies in Asia and the World.

27. Ali، I. 2012. 'Application of a Mining Algorithm to Finding Frequent Patterns in a Text Corpus: A Case Study of Arabic.' International Journal of Software Engineering and Its Applications، vol.6(3)، pp.127-134.

28. Alqurneh، A، A Mustapha، M Murad، and N Sharef. Forthcoming. 'Stylometric model for detecting oath expressions: A case study for Quranic texts.' Literary and Linguistic Computing journal.

29. Alrabiah، M، A Al-Salman، and E Atwell. 2013. 'The design and construction of the 50 million words KSUCCA King Saud University Corpus of Classical Arabic.' Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics.

30. Alrabiah، M، A Al-Salman، E Atwell، and N Alhelewh. 2014a. 'KSUCCA: A Key To Exploring Arabic Historical Linguistics.' International Journal of Computational Linguistics، vol. 5، pp.27-36.

31. Alrabiah، M، N Alhelewh، A Al-Salman، and E Atwell. 2014b. 'An Empirical Study On The Holy Quran Based On A Large Classical Arabic Corpus.' International Journal of Computational Linguistics، vol. 5، pp.1-13.

32. Alrehaili، S and E Atwell. 2013. 'Linguistics features to confirm the chronological order of the Quran.' Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics.

33. Alrehaili، S and E Atwell. 2014. 'Computational ontologies for semantic tagging of the Quran.' Proceedings of LRE-Rel 2: 2nd Workshop on Language Resource and Evaluation for Religious Texts.

34. Alruily، M. 2012. 'Using Text Mining to Identify Crime Patterns from Arabic Crime News Report Corpus.' PhD Thesis، De Montford University.

35. Attia، M، P Pecina، L Tounsi، A Toral، and J Van Genabith. 2011. 'Lexical Profiling for Arabic.' Proceedings of eLex'2011 Electronic Lexicography in the 21st Century.

36. Atwell، E، C Brierley، K Dukes، M Sawalha، and A Sharaf. 2011. 'An Artificial Intelligence Approach to Arabic and Islamic Content on the Internet.' Proceedings of NITS'2011 3rd National Information Technology Symposium، Riyadh.

37. Atwell، E، J Arshad، C Lai، L Nim، N Rezapour Asheghi، J Wang، and J Washtell. 2007. 'Which English dominates the World Wide Web، British or American?' Proceedings of CL›2007 Corpus Linguistics.

38. Atwell، E، K Dukes، A Sharaf، N Habash، B Louw، B Abu Shawar، A McEnery، W Zaghouani، and M El-Haj. 2010. 'Understanding the Quran: a new Grand Challenge for Computer Science and Artificial Intelligence.' Proceedings of GCCR›2010 Grand Challenges in Computing Research.

39. Atwell، E، L Al-Sulaiti، and S Sharoff. 2009. 'Arabic and Arab English in the Arab World.' Proceedings of CL2009 Corpus Linguistics.

40. Atwell، E، L Al-Sulaiti، S Al-Osaimi، and B Abu Shawar. 2004. 'A review of Arabic corpus analysis tools'، Proceedings of TALN'2004: Traitement Automatique des Langues Naturelles.

41. Atwell، E، N Abbas، B Abu Shawar، L Al-Sulaiti، A Roberts، and M Sawalha. 2008. 'Mapping Middle Eastern and North African Diasporas.' Proceedings of BRISMES'2008 British Society for Middle Eastern Studies.

42. Atwell، E. (ed.) 1993. 'Knowledge at Work in Universities - Proceedings of the second annual conference of the Higher Education Funding Council›s Knowledge Based Systems Initiative.' 146pp. Leeds University Press.

43. Atwell، E. 1982. LOB Corpus Tagging Project: Manual Postedit Handbook. Department of Linguistics and Modern English Language، University of Lancaster.

44. Atwell، E. 1993. 'The HEFC›s Knowledge Based Systems Initiative.' AISBQ: Artificial Intelligence and Simulation of Behaviour Quarterly، vol. 83، pp.29-34.

45. Atwell، E. 2008. 'Development of tag sets for part-of-speech tagging.' Ludeling A; Kyto M (ed.) Corpus Linguistics: An International Handbook، Volume 1، pp.501-526. Mouton de Gruyter.

46. Atwell، E. 2011. 'Exploiting New Technology and Innovation for Detecting Terrorist Activities.' Counter Terror Expo، London.

47. Bannister، A. 2014. ‘An Oral-Formulaic Study of the Quran.’ Lexington.

48. Baroni، M and S Bernardini. 2004. ‘BootCaT؛ Bootstrapping corpora and terms from the web.’ Proceedings of LREC’2004 Language Resources and Evaluation Conference.

49. Brierley، C، E Atwell، C Rowland، and J Anderson. 2013. ‘Semantic Pathways؛ a Novel Visualization of Varieties of English.’ ICAME Journal of the International Computer Archive of Modern English، vol. 37، pp.5-36.

50. Brierley، C، M Sawalha، and E Atwell. 2012a. ‘Boundary Annotated Qur›an Corpus for Arabic Phrase Break Prediction.’ Proceedings of IVACS’2012 Inter-Varietal Applied Corpus Studies.

51. Brierley، C، M Sawalha، and E Atwell. 2012b. ‘Open-source boundary-annotated corpus for Arabic speech and language processing.’ Proceedings of LREC’2012 Language Resources and Evaluation Conference.

52. Brierley، C، M Sawalha، and E Atwell. 2012c. ‘Visualisation of Prosody in English and Arabic Speech Corpora.’ Proceedinds of AVML’2012 Advances in Visual Methods for Linguistics.

53. Brierley، C، M Sawalha، B Heselwood، and E Atwell. forthcoming. ‘A verified Arabic-IPA mapping for Arabic tran-

scription technology، informed by Quranic recitation، traditional Arabic linguistics، and modern phonetics.' Journal of Semitic Studies.

54. Brockett A، E Atwell، O Taylor، and M Page. 1989. 'An Arabic text database and glossary system for students.' Proceedings of the Seminar on Bilingual Computing in Arabic and English.

55. Danso، S، E Atwell، O Johnson، A ten Asbroek، S Soromekun، K Edmond، C Hurt، L Hurt، C Zandoh، C Tawiah، J Fenty، S Etego، S Agyei، and B Kirkwood. 2013. 'A semantically annotated verbal autopsy corpus for automatic analysis of cause of death.' ICAME Journal of the International Computer Archive of Modern and Medieval English، vol. 37، pp.37-69.

56. Dukes، K and E Atwell. 2012. 'LAMP: a multimodal web platform for collaborative linguistic analysis.' Proceedings of LREC'2012 Language Resources and Evaluation Conference.

57. Dukes، K and N Habash. 2010. 'Morphological Annotation of Quranic Arabic.' Proceedings of LREC'2010 Language Resources and Evaluation Conference.

58. Dukes، K and T Buckwalter. 2010. 'A Dependency Treebank of the Quran using Traditional Arabic Grammar.' Proceedings of INFOS'2010 7th Informatics and Systems.

59. Dukes، K، E Atwell، and A Sharaf. 2010. ‘Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank.’ Proceedings of LREC›2010 Language Resources and Evaluation Conference.

60. Dukes، K، E Atwell، and N Habash. 2013. ‘Supervised collaboration for syntactic annotation of Quranic Arabic.’ Language Resources and Evaluation Journal، vol. 47، pp.33-62.

61. El-Beltagy، S، and A Ali. 2013. ‘Open issues in the sentiment analysis of Arabic social media: A case study.’ Proceedings of IIT’2013 Innovations in Information Technology.

62. El-Haj، M، U Kruschwitz، C Fox. Forthcoming. Creating language resources for under-resourced languages: methodologies، and experiments with Arabic. Language Resources and Evaluation journal.

63. Elarian، Yousef S. and Fayez M. Idris. 2010. ‹A Lexicon of Connected Components for Arabic Optical Text Recognition› First International Workshop on Frontiers in Arabic Handwriting Recognition، 22 August 2010. Istanbul.

64. Erradi، A، S Nahia، H Almerekhi، and L Al-kailani. 2012. ArabicTutor: a Multimedia m-Learning Platform for Learning Arabic Spelling and Vocabulary. Proceedings of ICMCS’2012 International Conference on Multimedia Computing and Systems.

65. Friginal، E and J Hardy. 2014. ‘Corpus-based Sociolinguistics: A Guide for Students.’ Routledge.

66. Garside، R and N Smith. 1997. ‘A hybrid grammatical tagger: CLAWS4.’ in Garside، R، G Leech and A McEnery (eds.) ‘Corpus Annotation: Linguistic Information from Computer Text Corpora.’ Longman، London، pp. 102-121.

67. Greene، B and G Rubin. 1971. ‘Automatic grammatical tagging of English.’ Technical report، Department of Linguistics، Brown University.

68. Hassan، H، N Daud، and E Atwell. 2010. ‘Connectives in the World Wide Arabic corpus.’ Proceedings of IVACS›2010 Inter-Varietal Applied Corpus Studies.

69. Hassan، H، N Daud، and E Atwell. 2013. ‘Connectives in the World Wide Web Arabic corpus.’ World Applied Sciences Journal (Special Issue of Studies in Language Teaching and Learning)، vol. 21، pp.67-72.

70. Jurafsky، D. and Martin، J. H. 2009. ‘Speech and Language Processing: An Introduction to Natural Language Processing، Speech Recognition، and Computational Linguistics’. New Jersey: Prentice Hall.

71. Karlsson، F، A Voutilainen، J Heikkila، and A Anttila (eds.). 1995. ‘Constraint Grammar: A Language-Independent System for Parsing Running Text.’ Mouton de Gruyter، Berlin and New York.

72. Kilgarriff، A. 2007. 'Re: [Corpora-List] history of corpus linguistics.' Corpora-List Archive، 6 January 2007.

73. Kilgarriff، A، V Baisa، J Bušta، M Jakubíček، V Kovář، J Michelfeit، P Rychlý، and V Suchomel. 2014. 'The Sketch Engine: ten years on.' Lexicography journal vol.1(1)، pp.1-30.

74. Kilgarriff، A، F Charalabopoulou، M Gavrilidou، J Jonannessen، S Khalil، S Johansson، R Lew، S Sharoff، R Vadlapudi، and E Volodina. 2013 'Corpus-based vocabulary lists for language learners for nine languages.' Proceedings of LREC'2013 Language Resources and Evaluation Conference.

75. Leech، G. 1992 ‹Corpora and theories of linguistic performance›، in Svartvik، J. ‹Directions in Corpus Linguistics›، pp 105-22. Mouton de Gruyter، Berlin.

76. Leech، G، R Garside، and E Atwell. 1983a. 'Recent developments in the use of computer corpora in English language research.' Transactions of the Philological Society، 1983، pp.23-40.

77. Leech، G، R Garside، and E Atwell. 1983b. 'The Automatic Grammatical Tagging of the LOB Corpus.' ICAME Journal: International Computer Archive of Modern and Medieval English Journal، vol. 7، pp.13-33.

78. Malmasi، S، and M Dras. 2014. Arabic Native Language Identification. In the proceedings of the EMNLP 2014

Workshop on Arabic Natural Language، 25 October 2014، Doha، Qatar.

79. McEnry، T. R Xiao and Y Tono. 2006. ‹Corpus-Based Language Studies: An advanced resource book›. Routledge، London & New York.

80. Merakchi، K and M Rogers. 2013 'The translation of culturally bound metaphors in the genre of popular science articles: A corpus-based case study from Scientific American translated into Arabic.' Intercultural Pragmatics journal، vol.10(2)، pp.341-372.

81. Mohammed، M and N Omar. 2011. 'Rule based shallow parser for Arabic language.' Journal of Computer Science، vol.7(10)، pp.1505-1514.

82. Mohit، B، A Rozovskaya، N Habash، W Zaghouani and O Obeid. 2014. The First QALB Shared Task on Automatic Text Correction for Arabic. In the proceedings of the EMNLP 2014 Workshop on Arabic Natural Language، 25 October 2014، Doha، Qatar.

83. Panju، M. 2014. 'Statistical Extraction and Visualization of Topics in the Quran Corpus.' MMath Thesis، University of Waterloo.

84. Rabiee، H. 2011. Adapting Standard Open-Source Resources To Tagging A Morphologically Rich Language: A Case Study With Arabic. Proceedings of RANLP'2011 Recent Advances in Natural Language Processing.

85. Roberts، A، L Al-Sulaiti، and E Atwell. 2005. 'aConCorde: towards a proper concordance of Arabic.' Proceedings of CL'2005 Corpus Linguistics.

86. Roberts، A، L Al-Sulaiti، and E Atwell. 2006 'aConCorde: Towards an open-source، extendable concordancer for Arabic.' Corpora journal، vol. 1، pp. 39-57.

87. Sawalha، M and E Atwell. 2008. 'Comparative evaluation of Arabic language morphological analysers and stemmers.' Proceedings of COLING'2008 Computational Linguistics.

88. Sawalha، M and E Atwell. 2009. 'Linguistically informed and corpus informed morphological analysis of Arabic.' Proceedings of CL'2009 Corpus Linguistics.

89. Sawalha، M and E Atwell. 2010a. 'Fine-Grain Morphological Analyzer and Part-of-Speech Tagger for Arabic Text.' Proceedings of LREC'2010 Language Resources and Evaluation Conference.

90. Sawalha، M and E Atwell. 2010b. 'Constructing and Using Broad-Coverage Lexical Resource for Enhancing Morphological Analysis of Arabic.' Proceedings of LREC'2010 Language Resources and Evaluation Conference.

91. Sawalha، M and E Atwell. 2011. 'Morphological analysis of classical and modern standard Arabic.' Proceedings OF ICCA'2011 International Computing Conference in Arabic.

92. Sawalha، M and E Atwell. 2012. ‘Visualization of Arabic Morphology.’ Proceedings of AVML’2012 Advances in Visual Methods for Linguistics.

93. Sawalha، M and E Atwell. 2013a. ‘Accelerating the processing of large corpora؛ using grid computing for lemmatizing the 176 million words Arabic Internet Corpus.’ Proceedings of WACL’2 2nd Workshop of Arabic Corpus Linguistics.

94. Sawalha، M and E Atwell. 2013b. ‘A standard tag set expounding traditional morphological features for Arabic language part-of-speech tagging.’ Word Structure journal، vol. 6، pp.43-99.

95. Sawalha، M and E Atwell. 2013c. ‘ Comparing morphological tag-sets for Arabic and English.’ Proceedings of CL’2013 Corpus Linguistics.

96. Sawalha، M، C Brierley، and E Atwell. 2012a. ‘Predicting phrase breaks in classical and modern standard Arabic text.’ Proceedings of LREC’2012 Language Resources and Evaluation Conference.

97. Sawalha، M، C Brierley، and E Atwell. 2012b. ’Prosody prediction for Arabic via the open-source boundary-annotated Qur›an corpus.’ Journal of Speech Sciences، vol. 2، pp.175-191.

98. Sawalha، M، C Brierley، and E Atwell. 2014. ‘Automatically generated، phonemic Arabic-IPA pronunciation tiers for the

boundary annotated Qur'an dataset for machine learning.' Proceedings of LRE-Rel'2، 2nd Workshop on Language Resource and Evaluation for Religious Texts.

99. Sawalha، M، E Atwell، and M Abushariah. 2013. 'SALMA: Standard Arabic Language Morphological Analysis.' Proceedings ICCSPA'2013 International Conference on Communications، Signal Processing، and their Applications، pp.1-6.

100. Sharaf، A and E Atwell. 2009. 'A Corpus-based Computational Model for Knowledge Representation of the Quran'، Proceedings of CL'2009 Corpus Linguistics.

101. Sharaf، A and E Atwell. 2012a. 'QurAna: Corpus of the Quran annotated with Pronominal Anaphora.' Proceedings of LREC'2012 Language Resources and Evaluation Conference.

102. Sharaf، A and E Atwell. 2012b. 'QurSim: A corpus for evaluation of relatedness in short texts.' Proceedings of LREC'2012 Language Resources and Evaluation Conference.

103. Sharoff، S. 2006. 'Open-source corpora: using the net to fish for linguistic data.' International Journal of Corpus Linguistics 11 (4)، pp. 435–62.

104. Tabrizi، A، and R Mahmud. 2013. 'Issues of coherence analysis on English translations of Quran.' Proceedings of ICC-

SPA'2013 International Conference on Communications، Signal Processing، and their Applications.

105. Wiechmann، D and S Fuhs. 2006. 'Concordance Software.' Corpus Linguistics and Linguistics Theory journal، vol.2، pp109-130

106. Yusof، R، R Zainuddin، M Baba، and Z Yusoff. 2010. 'Quranic words stemming.' Arabian Journal for Science and Engineering، vol.35(2)، pp.37-49.

قواعد البيانات الإلكترونية للمخطوطات التراثية العربية والإسلامية: الحاضر والمستقبل

د. سامح عويضة (*)
s.awaida@qu.edu.sa

(*) أستاذ مساعد في قسم هندسة الحاسوب في جامعة القصيم، المملكة العربية السعودية. حصل على درجة الدكتوراة في هندسة وعلوم الحاسب الآلي من جامعة الملك فهد للبترول والمعادن، والبكالوريوس والماجستير في الهندسة الكهربائية من جامعة هارتفورد، الولايات المتحدة الأمريكية. في السابق عمل محاضرًا في جامعة الملك فهد وجامعة الأميرة سمية. تشمل اهتماماته البحثية التعرف على الأنماط ومعالجة الصور والأنظمة المدمجة. بالإضافة إلى اثنتين من براءات الاختراع العالمية، نشر أكثر من خمسة عشر بحثًا علميًا في مجلات ومؤتمرات دولية.

ملخص

إنّ المشتغـل بالتراث العربي والإسلامي يلاحـظ نقصًا واسعًا في جهود رقمنة المخطوطـات، وندرة في الاستفادة من البرمجيـات والخوارزميات الحاسوبية ذات العلاقـة بمعالجة الوثائق التراثية؛ على النقيض من مدى استفادة العلوم الإنسانية عنـد الغرب لهذه التقنية. يذكر الباحث في هـذا المقال واقع رقمنة الوثائق التراثية من قواعد البيانات الإلكترونية لفهارس المخطوطات، وقواعد المخطوطات الرقمية، مـع ذكر أبرز المزايـا والعيوب. كما يقترح المواصفات المناسبـة لبناء قاعدة بيانات متنوعـة الصـور للمخطوطـات العربية والإسلاميـة الممسوحة ضوئيـا، مع تحديد المتطلبـات التي تمكن قاعدة البيانات المقترحـة من خدمة باحثي العلوم الإسلامية وعلـوم الحاسـب الآلي معًـا. تمّ تنظيـم قاعـدة البيانـات الإلكترونيـة المقترحـة للمخطوطات التراثية لتقوم وفق مهام محدّدة مسبقًا تساعد في تطوير البرمجيات الخاصّـة بباحثي الحاسب الآلي، وتمكينهم من خدمـة التراث العربي والإسلامي؛ والتي منها – على سبيل المثال – التعرف على الكتابة آليًا، والبحث عن صور الكلمات دون تعـرف، وتحليل صور الوثائق، والتعرف على الناسخ والتحقق منه، وغيرها من مستجدّات البحث العلمي.

١. مقدمة

إن العناية بتراث علمائنا السابقين، بحفظ أصول كتبهم الخطية من الواجبات الكفائيــة، ويشمــل ذلــك إيقــاف طلبة العلــم النبهاء عليهــا، واستخــدام الوسائل العصريــة التي توسع نطاق ذلك، وتذلّــل عقباته، كالبرامج الحديثة التي تعين على الوقوف على المعلومات بأسرع وقت، وعلى وجه فيه دقّة [١].

ويقّدر عــدد المخطوطات العربية والإسلامية اليــوم بثلاثة ملايين مخطوطة. تتطــرق العربية منهــا إلى موضوعات شتّــى، إذ تتسع دائرة عناوينهــا لتشمل كافة مجالات العلوم الإنسانية تقريبًا، بينما تحتوي المخطوطات الإسلامية على نصوص بعدّة لغــات – إضافةً إلى اللّغة العربية – كالفارسيــة والتركية العثمانية والأوزبكيّة وعّــدة لغــات أخــرى. هــذا، وتوجــد مجموعة ضخمــة مــن المخطوطــات والوثائق الإسلاميــة والتاريخية في المكتبات والمتاحف العالمية يجري حاليا مسحها وتخزينها رقميًــا – لأغراض الحفظ وتيســير الحصول عليها من قبــل الباحثين والمهتمين في جميــع أنحاء العــالم – من خلال المكتبات الرقمية المتوفرة علــى الشبكة العنكبوتية الإلكترونية (الإنترنت). لمــاّ كانت المخطوطات الأصلية بين يدي الباحثين ولا سيما مجاميــع المخطوطــات، كانــوا على مقــدرة من معرفة عمــر المخطوط بنــوع الورق والحبر، وتحديد اسم الناسخ الذي يكون مسجّلًا في آخر المجموع غالبًا، فلمّا توزعت رسائل المجموع بالتصوير الضوئي فقدنا أسماء الناسخ في عدد من الرسائل ما عدا الأخيرة مــن المجموع، مما يدعو إلى الحاجة إلى استخدام التقنية في التعرف على ناسخي هذه الرسائل.

تشّكل تقنيات المسح الضوئي والرقمنة للمخطوطات وسيلة وصول لكنوز التراث المخفيــة في مكتباتنــا. وقد اهتمّ حافظي زهير في مقالــه "دور تكنولوجيا المعلومات في حفــظ المخطوطات العربيــة" [٢] بذكر بعضـ التجارب الرائــدة في استخدام

النظـم الآليــة في معالجة المخطوطـات العربية وإتاحتها، مع ذكـر متطلبات رقمنة المخطوطـات وخصوصياتهـا ومواصفاتها. ومـع هذه الجهود المباركـة، فإنّ مجرد عرضـ صور من المخطوطة لا يعد كافيًا، لأنّ الصور غير قادرة على نقل العديد من الجوانـب الهامة للمخطوطة؛ بمـا في ذلك وصف المخطوط، ووصف ورقه واكتماله، ومحتـوى الصور، والكتابة والخط، وذكر تاريخ ومكان كتابة المخطوط بفهرس قابل للبحـث فيه، وغيرهــا من المعلومات الهامّة. لذا، فينبغـي ألّا يكون الهدف الرئيسي مـن رقمنة الوثائق التراثية مقتصرًا على توفير صور المخطوطات والوثائق فحسب، ولكـن على توفير وصف لمحتوى المخطوطات يتميز بالدقـة والشمولية، وكذلك على توفـير نصوصـ المخطوطـات لكي يتنسـى للباحثـين في المستقبل البحـث في صور المخطوطـات بسهولة ويسر، والعمـل على توفير الترجمة (الآليـة) لهذه النصوص لتذليل العلوم الإسلامية والعربية لسكّان العالم أجمع.

نظـرًا للتطورات الأخـيرة في علوم الحاسب الآلي، وخاصّـةً علم التعرف على الأنمـاط (Pattern Recognition)، فقـد أصبـح مـن الممكن لهـذه العلوم خدمة باحثـي التراث العربي والإسلامي بصورةٍ كبيرة. نذكر تاليًا أهمّ هذه العلوم وطرق إفادتها لباحثي العلوم الإنسانية:

أ. التعرف على النصوص المكتوبة آليًا (Automatic Text Recognition):

تقـوم هـذه البرامج بتحويـل صـور المخطوطات الممسوحـة ضوئيًـا إلى نص إلكـتروني ممّا يتيح البحث فيـه، والتعديل عليه، وطباعته ونشـره إلكترونيًا ضمن عـدد من العمليـات المفيدة. ولا شكّ أنّ برمجيات التعرف علـى النص المكتوب آليًا مـن أكـثر الأمور التي تفيد الباحثـين في العلوم الإنسانية، إذ تعـد بديلًا عن عملية النسخ والتفريغ اليدوي للمخطوط التي قد تستغرق الأيام أو الأشهر والسنين. وهي ميزة تفوق الوصف في تيسير علم البحث في المخطوطات، ولك أن تتصور كم في هذا مـن تذليل للعلوم إن توفّـر البحث في قاعدة بيانات تضم عـددًا ضخمًا من الصور

آليًــا، وخاصّةً إن استطاعت هذه العملية الآلية التفوق على نتائج التفريغ اليدوي في الدقّــة. وممّــا تعتمد عليه نتائج هــذه البرمجيات في نجاحها دقّــة نتائج برمجيات أخرى، مثل تحليل صور الوثائق ومعالجة اللغة الطبيعية.

ب. البحث عن صور الكلمات (Word Spotting):

كثـيرًا ما يحتــاج الباحثون في العلــوم الإنسانية إلى البحث عــن كلمة أو جملة محــددة في المخطــوط دون معرفة نصــ المخطــوط بأكملــه، أي دون أن تضطرّهم لقــراءة المخطوط بأكمله أو التعرف على نصــه آليًا للبحث عمّا يحتاجون. ولتطبيق هذه العملية نستخدم صورًا للكلمة أو الجملة المراد البحث عنها ونستخدم برمجيات علم التعرف على الأنماط للبحث عنها وتحديد مكانها في المخطوط.

ج. تحليل صور الوثائق (Document Image Analysis):

يهــدف علم تحليل صور الوثائق إلى تحليــل صور المخطوطات التراثية ووصف هــذه الصور للمساعدة في علوم أخرى. فعلى سبيل المثال، تستطيع برمجيات تحليل صــور الوثائق التعرف على مواقــع الفقرات الكتابية في الصفحة وتحديد إحداثيات محيــط هذه الفقــرات، كما تقوم بتقسيــم الفقرات إلى أسطر منفــردة لتجهيزها للتعــرف الآلي علــى الكتابــة، بالإضافــة إلى تحديــد الرسومات (مثــل الأختام أو الأشــكال) في صور الصفحات مــع وصفها وصفًا مفيدًا، والكثير من الأمور الأخرى التي تساعد برمجيات التعرف على الأنماط.

د. معالجة اللغة الطبيعية (Natural Language Processing):

يستخــدم علــم معالجة اللغــة الطبيعية قواعد اللغــة العربية ونتائــج التحليل الإحصائــي للنصوص الإسلامية في تمكين أجهزة الحاسب الآلي من فهم نصوص الوثائــق التراثية واستكشاف العلاقات والأنمــاط المفيدة، وهو من العلوم المساعدة المهمّة في علم التعرف على الأنماط. فعلى سبيل المثال، تُستخدم برمجيات معالجة اللغة الطبيعية في تحسين دقّة نتائج التعرف على الكتابة آليًا.

٥. التعرف على الناسخ والتحقق منه (Writer Identification and Verification):

تعنى عملية التعرف على الناسخ أو الخاطّ بتحديده آليًا من مجموعة من النسّاخ من خلال تحليل عينات من صور خطوطهم ضمن قاعدة بيانات لخطوط النسّاخ. كما تعني عملية التحقق من الناسخ التأكيد الآلي لصحة نسبة المخطوط لناسخ معيّن من خلال مقارنتها مع عينات سابقة لصور خطّه. ممّا قد يعطي قيمة علمية للمخطوط كأن يكون الناسخ ممن عُرف بفن معين فيسهل على الباحث التعامل مع المخطوط في فهم النص وتوجيه عبارته من خلال ما عُرف من منهج الناسخ في منسوخاته أو منهج المؤلف في مؤلفاته السابقة، وتزيد أهمية هذا البند في الاهتداء لاسم الناسخ أو المؤلف فيما عُثر عليه من مخطوطات ناقصة الأول والآخر؛ فيتم التعرف عليه من خلال وجود قاعدة ضخمة من صور المخطوطات — التي عرف اسم ناسخها أو مؤلفها - تعرض على إثر بعضها فيظهر اسم الناسخ أو المؤلف، وقد يهتم هذا العلم - بدلًا من تحديد اسم الناسخ - بتحديد عصر أو قرن المخطوط، أو نوع الخط المستخدم في الكتابة، ممّا يساعد في تثمين قيمة المخطوطات الأثرية وكشف المزيف منها.

إنّ أحدث التقنيات حاليًا لا يمكنها التعامل مع المخطوطات والوثائق التاريخية في البحث عن النصوص، أو معرفة المحتوى، فضلًا عن ترجمته بصورة ناجحة. لذا فلا بدّ من بناء قاعدة بيانات رقمية ضخمة للباحثين تحتوي على الوصف المناسب لكل مخطوط بالإضافة إلى نصوص المخطوط، مع تجهيز الأدوات المناسبة للباحثين في مجال علوم الحاسب الآلي والتي تساعدهم في بناء البرمجيات الخاصة بالبحث والتحليل والترجمة والتعرف على الأنماط آليًا.

نقوم بتقسيم ما تبقّى من هذا المقال على النحو التالي؛ نناقش في الفصل الثاني واقع فهارس المخطوطات الرقمية مع ذكر أبرز المؤلفات والمواقع المعنية

بفهارس المخطوطات العربية والإسلامية، كما نخصص الفصل الثالث لذكر جهود الباحثين والمؤسسات في قواعد بيانات المخطوطات الرقمية، مع إيراد أبرز هذه القواعد وميزاتها وعيوبها، ونذكر في الفصل الرابع الجهود المبذولة في محاولة جمع قواعد بيانات خاصّة بخطوط العلماء السابقين وناسخي المخطوطات، ثمّ نقدم في الفصل الخامس مقترحًا لمواصفات قاعدة بيانات مستقبلية للمخطوطات الإسلامية والعربية وعرضها بالتقنية الرقمية، مع الخيارات والميزّات التي تمكن قاعدة البيانات المقترحة من خدمة باحثي العلوم الإسلامية وعلوم الحاسب الآلي معًا، وأخيرًا، يتم عرض النتائج والتوصيات في الخاتمة في الفصل السادس.

٢- فهارس المخطوطات الرقمية

يُعدُّ فن فهرسة المخطوطات علمًا قائمًا بذاته، إذ يعنى بضبط المخطوطات وتوثيقها، ثم تسجيلها في قوائم، أو بطاقات، أو كتب لحفظها. فيقوم المفهرس بوضع كل ما يتعلق بالمخطوط من معلومات في هذه الفهارس، كعنوان المخطوط، واسم المؤلف، واسم الناسخ، وتاريخ النسخ، وعدد الأوراق، وعدد السطور في كل صفحة، وحجم الورق، ونوع الورق، ونوع الحبر، وحال المخطوط، كما أنَّه يضع اقتباسًا من بداية المخطوط ونهايته، وغير ذلك من أوصاف يراها المفهرس، وهي تختلف من مكتبة إلى أخرى. وقد ذكر صلاح الدين المنجد في كتابه "قواعد فهرسة المخطوطات العربية" [٣] تاريخ فهرسة المخطوطات العربية والإسلامية، بالإضافة إلى بعض القواعد والأساليب المستخدمة في فهرسة المخطوطات.

ومع عدم توفر مصادر رسمية لعدد فهارس المخطوطات في العالم، إلا إنَّه يمكننا تقدير عددها بالآلاف، بل قد يصل إلى عشرات الآلاف؛ بدليل أنَّ مؤسسة الفرقان جمعت في كتابها "المخطوطات الإسلامية في العالم" ما يزيد على خمسة عشر ألف فهرس للمخطوطات الإسلامية [٤]. كذلك الفهارس التي جمعتها مؤسسة آل البيت للفكر الإسلامي في فهرسها الشامل وصلت إلى ألف وستمائة

فهرسٍ للمخطوطات [٥] مع أنها لم تستوعب جميع العلوم الإسلامية. هذا بالإضافة إلى العديد من المكتبات الخاصة والشخصية التي لم يوقف على فهارس مخطوطاتها.

جهود الباحثين المنشورة في جمع فهارس المخطوطات العربية والإسلامية

تزخر المكتبات بمخطوطات التراث العربي والإسلامي في مختلف بقاع العالم، لذا فإنَّك لا تكاد تزور دولة من هذه الدول إلا وتجد فيها فهرسًا أو أكثر لمخطوطات مكتباتها، إمَّا باللغة العربية وإمَّا بلغة الدولة الأمّ. وقد قام الباحثون على مدى التاريخ بعدَّة محاولات لجمع هذه الفهارس؛ وذلك للتعريف بها وتسهيل عمل الباحثين والمهتمين بهذا العلم، نذكر على سبيل المثال كتاب "تاريخ الأدب العربي" [٦]، لمؤلفه المستشرق الألماني كارل بروكلمان، الذي صدر الجزء الأساسي منه في مجلدين عامي ١٨٩٨ و١٩٠٢ ثم تلته المجلدات الثلاثة الكبيرة أعوام: ١٩٣٧ و١٩٣٨ و١٩٤٢ م. وقد يتصور الباحث للوهلة الأولى أنَّ هذه الموسوعة كتاب في تاريخ الأدب، بينما هي في الحقيقة سجل للمصنفات العربية المخطوط منها والمطبوع. ومن هذه الجهود أيضًا كتاب "تاريخ آداب اللغة العربية" لجورجي زيدان [٧]، و"تاريخ التراث العربي" لفؤاد سزكين [٨] الذي طبع باللغة الألمانية، ثم تولت جامعة الإمام محمد بن سعود ترجمته إلى العربية، وصدر منه عشرة أجزاء. كما قام السيد رزق الطويل في كتابه "مقدمة في أصول البحث العلمي وتحقيق التراث" [٩] بتكريس قسم منه لذكر فهارس المخطوطات الإسلامية. وبالجملة، فإنّ المؤلفات في هذا الباب كثيرة، وقد تصل إلى خمسين مؤلفٍ وزيادة، نذكر في العناوين التالية أبرزها.

فهارس "معهد المخطوطات العربية"

يُعد معهد المخطوطات العربية واحدًا من أقدم المراكز المعنيّة بالمخطوطات في الوطن العربي. تأسس المعهد سنة ١٩٤٦ وهو ملحق بالمنظمة العربية للتربية

والثقافـة والعلوم مع تمتعه بشخصية معنويـة مستقلّة. يُعنى المعهد بالتراث العربي المخطـوط بمختلف أصعدته؛ جمعًا، وإتاحـةً، وصيانةً، وترميمًا، وفهرسةً، وتعريفًا، ودراسـةً، وتوظيفًـا. للمعهـد عدد مـن الإصـدارات المهتمة بفهارسـ المخطوطات العربيـة، نذكـر منها "فهارسـ المخطوطات العربيـة في العـالم"، و"دليل مكتبات المخطوطـات في الوطن العربـي"، وعدد من فهارس المخطوطات في دول العالم مثل "فهرس مخطوطـات جامعة الإسكندرية" وغيرها، بالإضافـة إلى مؤلفات فيصل الحفيان – مدير معهد المخطوطات العربية –.

ولا شكّ أنّ من أهمّ مؤلفات المعهد "فهرس المخطوطات المصّورة" [١٠]، وهو فهرس للمخطوطات المصوّرة في معهد المخطوطات العربية ومقسّم حسب الفن من تاريـخ، وعلـوم، وعارف عامّة، وسـيرة نبوية، وأدب، وفقه، وحديـث، وتفسير. وقدّ تمت طباعة هذا الفهرس على عدّة مراحل زمنية تجاوزت خمسة عقود من الزمن، إذ تمّ طباعة المجلد الأول في عام ١٩٥٤ م، وثمّ طبع آخر مجلد عام ٢٠١١ م. وممّا يعيب مطبوعات المركز الانقطاع الطويل في أعمالهم، وعدم الاستمرارية في الجهود المبذولـة. كما أنّ المعهد من أوائل الذين تبنّوا سياسة رقمنة المخطوطات والفهارس منـذ عشـرات السنين، ولكنّنا لا نجد لهم برنامجًا متميـزًا من إصداراتهم في هذا الباب حتى يومنا هذا.

وللفائدة، فقـد نشر المعهد موقعًا باسم "خزانة المخطوطات" [١١]، يذكرون فيـه القدرة على البحث الرقمي عن المخطوطات التابعـة للمعهد، وهو قيد التعديل والتجريب منذ فترة، ونتائجه غير مرضية. كما أنّ موقع المعهد على الإنترنت مليء بالروابط التي لا تعمل والمعلومات الناقصة.

"المخطوطات الإسلامية في العالم" لمؤسسة الفرقان

قامـت مؤسسة الفرقان للـتراث الإسلامي [١٢] بعمل مسـح وإحصاء شامل لمجموعـات المخطوطات الإسلامية في أنحاء العالم شملت فهارس مئة وستة بلدان،

ثم نشرت نتائج هذه الدراسة في كتاب "المخطوطات الإسلامية في العالم" بنسختيه الإنجليزية [١٣] والعربية [٤]. وتذكر الموسوعة دراسات حول عدد المخطوطات، واللغات التي كتبت بها، كاللغة العربية والفارسية والتركية والأوردوية والسواحيلية؛ وأماكن تواجد تلك المخطوطات، كما تناولت الدراسة المجموعات التي فُهرست، وعناوين الفهارس التي تم نشرها، والتي لم تصدر بعد في تلك المكتبات، إضافة إلى ذكر بعض عناوين نوادر المخطوطات. وقد ذكرت مؤسسة الفرقان أنَّها قامت بجمع هذا الكتاب مما يزيد على خمسة عشر ألف فهرس للمخطوطات الإسلامية. مع الإشارة إلى أن كتاب "المخطوطات الإسلامية في العالم" هو أقرب لموسوعة تقوم بالتعريف بمكتبات العالم التي تحتفظ بفهارس للمخطوطات الإسلامية مع التعريف بهذه الفهارس، من كونها مجموعًا لفهارس المخطوطات كما هو حال "فهارس آل البيت" التي نناقشها تاليًا – مثلًا –.

فهارس آل البيت

ومن المشاريع المؤسسية المهتمة بحصر المخطوطات العربية الإسلامية استنادًا على الفهارس المطبوعة والمخطوطة للمكتبات العامة والخاصة، مشروع "الفهرس الشامل للتراث العربي الإسلامي المخطوط" الذي سبقت الإشارة إليه، والمعروف باسم "فهارس آل البيت" [٥] إذ قامت بجمعه مؤسسة آل البيت للفكر الإسلامي، فخرج في خمسة وعشرين مجلد. وقد اعتمدت المؤسسة في تحقيق هذا العمل على ما جمعته من الفهارس والمصادر والدراسات المختصة التي جاوزت الألف وستمائة فهرسٍ للمخطوطات في سبع عشرة لغة. وقد قامت المؤسسة بطباعة المشروع على مراحل عدّة حسب الموضوع، فطبعت فهرس مخطوطات التفسير وعلومه في مجلدين، وفهرس المصاحف المخطوطة ومخطوطات رسم المصحف في مجلد، وفهرس مخطوطات الحديث النبوي الشريف وعلومه ورجاله في ثلاثة مجلدات، وفهرس المخطوطات العربية في مكتبة تشستربيتي في إيرلندا في ثلاثة مجلدات، وفهرس مخطوطات التجويد في مجلد، وفهرس مخطوطات القراءات في مجلد،

وفهرس مخطوطات السيرة والمدائح النبوية في مجلدين، وفهارس مخطوطات الفقه وأصوله في اثنتي عشرة مجلد. ومع ضخامة عمل مؤسسة آل البيت في فهارسها، إلّا إنَّــه لا يـزال ناقصًا فلم يشمل جميع العلوم بعد، بل إنَّه متوقف حاليًا كما علمتُ من المؤسسة نفسها، وآخر هذه المجلدات من الفهرس طبع قبل عشر سنين في عام ٢٠٠٤ م. ومـن أبرز المآخذ على هذا الفهرس أنَّه لم يعتمد على مشاهدة المخطوطات التي تمـت فهرستها؛ بل اقتصر على النقل من الفهارس الأخرى؛ ممّا أدى إلى استمرار بعض الأخطاء كنسبة بعض المخطوطات إلى غير أصحابها – مثلًا – اعتمادًا على ما ذكر في الفهارس الأخرى.

وعلـى الرغم من عدم فهرسـة بعض المكتبات لمخطوطاتهـا أو عدم نشر هذه الفهارس، فإن الكثير من مكتبات العالم التي تحـوي مخطوطات عربية وإسلامية صنعـت فهرسًـا ورقيًـا مطبوعًا لهـذه المخطوطـات، لكـنَّ ظاهرة رقمنـة فهارس المخطوطـات لا تزال محدودة جـدًا؛ فالمكتبات ما زالت تعتمد على طباعة الفهارس ورقيًـا، مما يجبر الباحثين على البحث اليدوي في هـذه الفهارس، فيستهلك منهم الكثـير مـن الوقت والجهد، عدا عن ندرة بعض هـذه الفهارس المطبوعة لقدمها أو نفـاد المطبـوع منها. ولذا قام عدد مـن المؤسسات بمحاولة جمع هـذه الفهارس في مطبـوع موحّد يسهّل على المهتمين عملية البحث، وسنحـاول في هذا القسم التعريف بأشهـر فهارسـ المخطوطات الرقميـة. وبالرغم مـن نشر هذه المجاميـع لفهارس المخطوطـات، فإنّ خروج غالـب هذه المحاولات بصورة ورقية لا يـزال يشكل عائقًا في عملية البحث العلمي.

الجمع اليدوي لفهارس المخطوطات

من جهـود رقمنة فهارس المخطوطات، قيام بعضـ المؤسسات والمواقع المهتمة بالـتراث العربي والإسلامـي بجمع فهارس المخطوطات المطبوعـة ومسحها ضوئيًا ووضعهـا علـى صورة ملفـات (PDF أو Word) علـى الشبكة الإلكترونيـة. ومع أنَّ

غالـب هذه الملفات هي صور لصفحات هـذه الفهارس، مما يعني عدم القدرة على البحـث إلكترونيًـا فيها، إلا إنَّ بعضها قد تم تفريغه إلى نصّ إلكتروني، إمَّا آليًا أو يدويًا أو آليًا قد تمّت مراجعته يدويًا. إنَّ هذه الخطوات – وإن كانت أولية في رقمنة الفهارس – إلا إنَّها مهمّة، خاصة في الفهارس المطبوعة القديمة وضرورية لرقمنة فهارس المخطوطات كليًّا، ولن تكون تامة حتى يتم تفريغها كلها، وتدقيقها، ووضعها في قواعـد بيانـات منظمة. نذكـر على سبيل المثـال فهارس المخطوطـات في الموقع الإلكتروني لـ"مركـز ودود للفهارس وكتب التحقيـق" [٢٤]، وفهارس المخطوطات المجموعـة في برنامـج "المكتبـة الشاملـة" [١٤]، وموقـع "جامـع المخطوطـات الإسلامية" [١٥]، وبعض المنتديات الإلكترونية المهتمة بالعلوم الشرعية [١٦].

قواعد فهارس المخطوطات الرقمية

وبعـد أنْ وقفنـا علـى أهم الجهـود في جمع فهارسـ المخطوطـات الإسلامية المطبوعـة وتصويرها ضوئيا، نذكر هنا قواعد فهارسـ المخطوطات الرقمية، وهي خطـوة أساسيـة في رقمنة المخطوطـات، إذ تسمح هذه القواعـد بالبحث في حقول بطاقـة كل مخطـوط على حدة، مثـل اسم المؤلف وعنـوان المخطـوط وسنة النشر وغيرهـا من المعلومـات، إلا إنَّ هـذه الفهارس تفتقر عادة إلى خدمـة تصفح صور المخطوطات أو تحميلها.

مكتبة الفرقان الرقمية

من هذه القواعد نجد مكتبة الفرقان الرقمية [١٧] إذ يسمح الموقع الإلكتروني بالبحث في المخطوطـات. إلا إنّ واجهة البحث حاليًا تعمل باللغة الإنجليزية (يذكر الموقع أن الواجهة العربية تحت التطوير) ولا شك أنَّ الموقع بحاجة إلى إضافة الكثير من المخطوطات ليصبـح مرجعًا معتبرًا في البحث الرقمي في فهارس المخطوطات، إذ يذكـر الموقع أنَّه يحوي فهارس ٢٧٠٥٥ مخطوطًـا، وهذا العدد في ازدياد. يحوي الموقع الكثير من الخيارات التي تمكن الباحثين من البحث حسب المؤلف، والموضوع،

والمكتبــة، ولغــة المخطوط، وغيرها مــن الخيارات المفيدة. كما تُظهــر نتائج البحث معلومات المخطوط بطريقة منظمة، مع فرز كل جزء من بطاقة المخطوط (العنوان، اســم المؤلف، اسم الناســخ، ...) بحقل منفرد، مع إمكانية ربط بعض هذه الحقول إلكترونيًا، كما يظهر في رسم توضيحي ١.



Manuscript Details | Read Catalogue Introduction

Country:	Algeria
City:	Bejaia
Library:	Private library of Shaikh Lmuhub Ulahbib
Catalogue:	Catalogue of Islamic Manuscripts in the private library of Sheikh Lmūhūb Ulaḥbīb, Bejaia - Algeria Prepared by Djamel-Eddine Mechehed, Edited by Ayman Fuad Sayyid

MANUSCRIPT DETAILS

Manuscript Id	SP 18407
Title	[كتاب* في الفقه]
Author	محمد الموهوب بن البشير بن لحبيب، 1237هـ/1821م؛ تاريخ وفاة المؤلف الميلادي 1899.
Subject	الفقه.
Date of Authorship	1298هـ/1880م؛ تاريخ التأليف الميلادي 1881
Number of folios	27
Lines	25
Size	17cm × 21cm
Written Surface	11cm × 17cm
Script Type	Maghribī
Ink colour	أسود، الأحمر
Language	Arabic
Reference	La bibliothéque savante de cheikh al Muhub ; lettrés locaux et culture écrite au kabylie au milieu duXIX éme siécle. A paraître (publisud,Paris)\arsles manuscrits de botanique et médcine en kabylie au XIX siecle, ANNALI, instito universitario orientale, fasicolo 1-4. Napoli 1999، ج 59، ص 78.98؛ ANNALI,les manuscrits de botanique et médcine en kabylie au XIX siecle,Napoli 1999، ج 59، ص 69-92.
Condition	سيئة، وبها بتر من أولها وآخرها.
Notes & Comments	مكان التأليف : قرية تالة وزرار أيث ورثيلان – منطقة القبائل – الجزائر. العناوين بالأحمر.
Library	الشيخ الموهوب أولحبيب الخاصة/بجاية
Classmark	F0010

رسم توضيحي ١ : صورة توضيحية لنظام البحث في مكتبة الفرقان الرقمية.

موقع يوسف زيدان الإلكتروني

يقـوم موقـع يوسف زيـدان الإلكـتروني [١٨] بتوفير قاعدة بيانـات لفهارس المخطوطـات ، ويتيح للمتصفح البحث فيها ، ويحوي هـذا الموقع مجموعة قليلة من فهارسـ المخطوطـات العربية والإسلاميـة، وهي: فهرس مكتبـة رفاعة الطهطاوي (قرابـة ١٨٠٠ مخطوطة) ، وفهرس المخطوطات العلميـة بمكتبة بلدية الإسكندرية (٤٧٠ مخطوطـة) ، وفهرس مخطوطـات شبين الكـوم (٣٥٠ مخطوطة) ، وفهرس مخطوطـات الفقـه بمسجـد أبـى العباسـ المرسـى (٨٠٠ مخطوطـة) ، وفهرسـ مخطوطات المعهـد الديني بسُمُّوحَـةِ (٢٠٤ مخطوطة) ، وفهرسـ مخطوطات دار الكتـب بطنطا (٤٤٧ مخطوطة). ويظهر رسم توضيحي ٢ نظام البحث في فهارس المخطوطات في موقع يوسف زيدان.

فهارس المكتبات الخطية

بحـــث متقدم

إشارة :

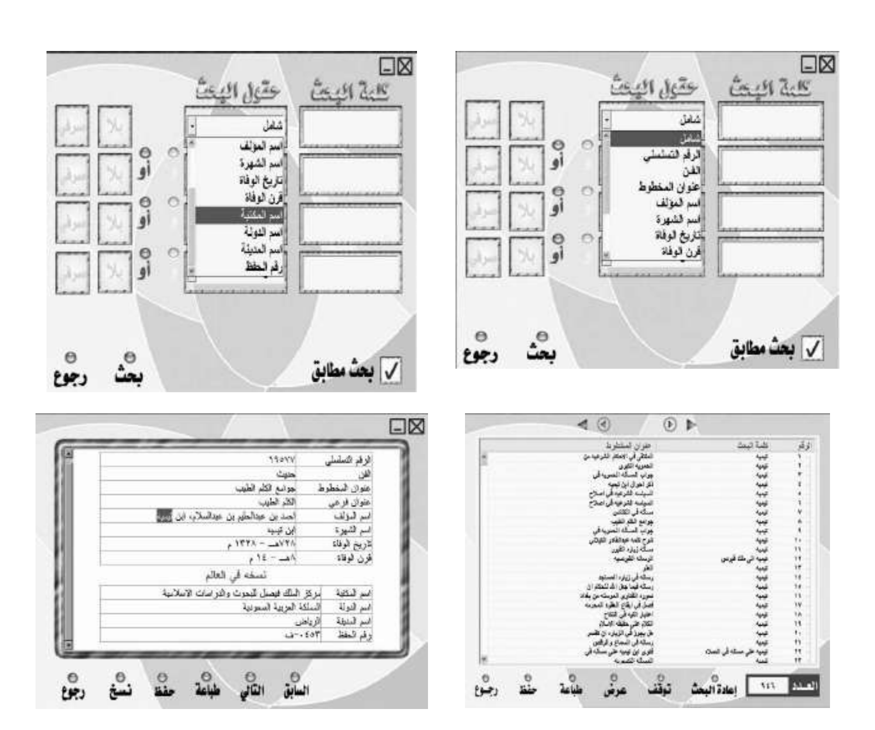
يشتمل هذا الباب من الموقع ، على فهارس وصفية كاملة لمكتبات خطية مهمة .
ويمكن البحث فى محتوى هذا الباب ، بكلمة من عنوان المخطوطة ، أو كلمة من اسم مؤلفها ، أوكلمة من بدايتها أو موضوع المخطوطة أو رقم حفظها.

حدّد الفهرس	جميع الفهارس
العنوان	
المؤلف	
الموضوع	
البداية	
رقم الحفظ	
	بحث

رسم توضيحي ٢ : صورة توضيحية لنظام البحث في موقع يوسف زيدان.

"خزانة التراث" لمركز الملك فيصل

كمـا قام مركز الملك فيصل للبحوث والدراسات الإسلامية بالرياض في العام ١٩٩٢ م بإنشـاء قاعـدة بيانات باسم "خزانة الـتراث" [١٩] تشتمل على فهارس المخطوطـات الإسلاميـة في جميـع المكتبـات والخزانـات ومراكـز المخطوطات في العالم، حتى بلغ عدد عناوين هذه المخطوطات أكثر من مائة وعشرين ألف. تتضمن خزانـة الـتراث فهارس مخطوطـات عربية وإسلاميـة من مختلف أنحـاء العالم، بالإضافة إلى فهارس لمخطوطات مكتوبة بلغات أجنبية ترجمها المركز إلى العربية. كمـا يوفر البرنامج مداخل بحث متعـددة يستطيع الباحث من خلالها الوصول إلى المعلومـة التي يريدها بكل يسر وسهولة، مثل البحث في الفن، وفي عنوان المخطوط، والعنـوان الفرعـي، وفي اسم المؤلف، واسـم الشهرة، وتاريخ الوفـاة، وقرن الوفاة، واسـم المكتبة، واسم الدولـة، واسم المدينة، ورقم الحفـظ. ويظهر رسم توضيحي ٣ عـدّة صـور لشاشات برنامج "خزانة التراث". ومن أبـرز عيوب برنامج "خزانة الـتراث" - مـع تميزه في مجاله - أنَّه لا يعمـل إلاّ على نظام تشغيل الويندوز فقط، وأنّه لم يتم تحديثه منذ سنين، مما يحدّ من استخدام البرنامج على أكثر الأجهزة الحديثة.



رسم توضيحي ٣ : عدّة صور توضيحية لشاشات برنامج "خزانة التراث".

"خزانة الماجد للتراث" لمركز جمعة الماجد

ومن البرامج المميزة في فهارس المخطوطات برنامج "خزانة الماجد للتراث" [٢٠] المصمم من قبل شركة الدار العربية لتقنية المعلومات للتعريف بمخطوطات مركز جمعة الماجد للثقافة والتراث والبحث فيها. يحوي البرنامج أشمل قاعدة بيانات إلكترونية لمخطوطات مركز جمعة الماجد. تذكر بطاقة المخطوط عنوان المخطوطة، والعنوان البديل، والمؤلف، والناسخ، وبدايتها ونهايتها، والملاحظات المتعلقة بالمخطوطة، وعدد الأوراق، واللغة، ويظهر رسم توضيحي ٤ مثالًا لإحدى بطاقات المخطوطات في البرنامج. ويقدم البرنامج خدمة البحث عن أي مخطوطة وذلك بالإشارة إلى ما يراد البحث فيه من العنوان أو العنوان البديل أو بداية المخطوط أو نهايته أو عن طريق مؤلفه أو ناسخه أو جميع ما سبق، مما يتيح للباحث

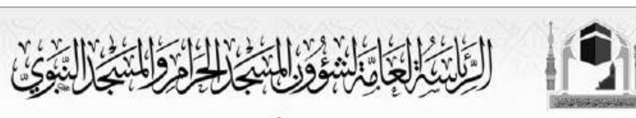
سرعة الوصول للمخطوطة المطلوبة. إلا إنَّ هذا البرنامج كسابقه -برنامج "خزانة الـتراث" لمركز الملـك فيصل - يعتمد علـى نظام تشغيل معيّن بـدلًا من جعله على الشبكة العنكبوتية العالمية ممّا يحدّ فعليًا من فائدته؛ لغياب التحديثات له.

رقم المادة	243171
العنوان الرئيسي	تدريب المبتدي وتهذيب المنتهي
المؤلف	البلقيني، عمر بن رسلان بن نصير الكناني العسقلاني الشافعي سراج الدين، أبو حفص، ت 805هـ
عدد الأوراق	16ق
لغة/ لغات المادة	العربية

رسم توضيحي ٤ : صورة توضيحية لبطاقة مخطوط في برنامج "خزانة الماجد للتراث".

فهارس مخطوطات مكتبة المسجد النبوي

ومـن الجهـود الحديثة في رقمنة فهارسـ المخطوطات جهـود الرئاسة العامّة لشؤون المسجد الحرام والمسجد النبوي، إذ قاموا برقمنة فهارس مخطوطات مكتبة المسجد النبـوي وتوفيرها للباحثين على شبكة الإنترنت [٢١]. تقوم المؤسسة دوريًا بإضافـة بطاقات للمخطوطـات، وفي زيارة حديثة للموقـع تم حصر فهارس ٦٢٢٠ مخطوطًـا. يسمح الموقع بالبحث في الفن، وفي عنـوان المخطوط، وفي أول المخطوط وآخره، أو في جميع بيانات المخطوط السابقة. ويظهر رسم توضيحي ٥ مثالًا للبحث في موقـع مكتبـة المسجد النبوي. يفتقد الموقع إلى تنوع مجـالات البحث، كما تفتقد بطاقـات المخطوطات إلى الشمول، وهناك نقص في كثير منها. إلا إنَّ الاستمرار في تحديـث هذا العمل وتحسينه قد يجعله من المراجع الرئيسة في فهارس المخطوطات الرقمية في المستقبل.



رقم الحاسب	التصنيف	...وط	المؤلف	
8970	الحديث	...جه (جزء منها) .	ابن المرزبان: أبو عبد الله، الحسين بن أحمد بن عبد الله الفارسي . ت:409هـ	عرض البيانات
8971	تاريخ وتراجم وأنساب	...ابن مندة(مرتب على ...م) .	الخلال: أبو عبد الله، الحسين بن عبد الملك بن الحسين الخلال البحرى . ت:532هـ	عرض البيانات
8972	الحديث	...اعة من الصحابة	ابن أبي عزرة: أبو عمر، أحمد بن حازم بن محمد الغفاري المكي . ت:276هـ	عرض البيانات
8973	الحديث	كتاب الجمعة وفضلها	المروزي:أبو بكر أحمد بن علي بن سعيد . ت:292هـ	عرض البيانات
8974	وعظ وفضائل أعمال	أبواب السعادة في أسباب الشهادة	السيوطي جلال الدين أبو الفضل عبد الرحمن بن أبي بكر بن محمد الخضيري - ت : 911هـ	عرض البيانات
8975	العقيدة	تشييد الأركان في ليس في الامكان أبدع مما كان	السيوطي جلال الدين أبو الفضل عبد الرحمن بن أبي بكر بن محمد الخضيري - ت : 911هـ	عرض البيانات
8976	السيرة	ريح النسرين فيمن عاش من الصحابة مائة وعشرين	السيوطي جلال الدين أبو الفضل عبد الرحمن بن أبي بكر بن محمد الخضيري - ت : 911هـ	عرض البيانات
8977	العقيدة	لبس اليلب في الجواب عن إيراد أهل حلب	السيوطي جلال الدين أبو الفضل عبد الرحمن بن أبي بكر بن محمد الخضيري - ت : 911هـ	عرض البيانات
8978	الأدب	المقامة المصرية في التصوف	السيوطي جلال الدين أبو الفضل عبد الرحمن بن أبي بكر بن محمد الخضيري - ت : 911هـ	عرض البيانات
63189	علوم الحديث	تاريخ ابن أبي خيثمة	ابن أبي خيثمة:أبو بكر أحمد بن زهير بن حرب النسائي البغدادي - ت: 279هـ	عرض البيانات

... 613 614 615 616 617 618 619 620 621 622

رقم الحاسب: 8970

أول المخطوط	آخر المخطوط
وحسبنا الله ونعم الوكيل . أخبرنا الحافظ أبو طاهر ... أحمد بن محمد السلفي الأصبهاني قال: أخبرنا ... أبو عبد الله القاسم بن الفضل ...	وقوله الضرب بين يديك بالسيف دليل على المحبة لأن من أحب حبا أثاب عنه مجاهد أحد ه وذب عنه ...

رقم الحاسب: 8970

ناسخ المخطوط	سنة النسخ	درجة الخط	عدد الأوراق	طول	عرض	عدد الأسطر	نوع الخط
-	—		8	42	30	19	الرقعي

رسم توضيحي ٥ : مثال لخدمة البحث بفهارس مخطوطات المسجد النبوي.

وللقارئ مراجعة موقع جامعة ميتشغان الأمريكية للمزيد من المؤلفات والمواقع المعنيـة بفهارس المخطوطات والإسلاميـة، إذ قاموا بفهرسـة مجموعة كبيرة من روابـط فهارس المخطوطـات الإسلاميـة الرقمية حـول العالم[٢٢] مـع توصيف مختصر لها. ونحن سنخصص الفصل القادم لذكر جهود الباحثين والمؤسسات في قواعـد بيانات المخطوطات الرقمية، مع إيراد أبرز هذه القواعد وميزاتها وعيوبها. ومن الجدير بالذكر أنَّ قواعد البيانات الإلكترونية للمخطوطات تعد كذلك فهرسًا لهـذه المخطوطـات، فمـن الممكن إضافـة كل ما سيذكـر في الفصل القـادم ضمن فهارس المخطوطات الرقمية في هذا الفصل.

٣. قواعد المخطوطات الرقمية

اختلف الباحثون في تقدير عدد فهارس المخطوطات العربية والإسلامية، كما اختلفوا في تقدير عدد المخطوطات نفسها. ويرجع هذا الاختلاف إلى عدّة أسباب، أهمّها عدم وجود مؤسسة رسمية تعنى بإحصاء دقيق ودوري لهذه المخطوطات. كذلك من الأسباب عدم الاتفاق في تحديد ما يراد إحصاؤه، فهل يراد حصر المخطوطات الإسلامية بلغاتها المختلفة أو إدخال المخطوطات العربية بعلومها كافّة؟ وأيضًا الاختلاف في تحديد القرون المراد إحصاء مخطوطاتها، ووجود عدد من المخطوطات ذات الملكية الخاصّة أو المخطوطات المفقودة أو المسروقة التي لا يمكن إحصاؤها إلى غيرها من الأسباب التي تجعل التفاوت -حتى في الأرقام التقديرية للمخطوطات- كبيرًا بين المصادر المختلفة. فمنهم من يقدّرها بمليوني مخطوطة، أو ثلاثة ملايين، ومنهم من يوصل عددها إلى خمسة ملايين مخطوطة، مع شبه اجماعهم على أنَّ عدد المخطوطات العربية والإسلامية بالملايين. ويذكر الباحثون أنَّ أكثر هذه المخطوطات في تركيا (٢٠٠،٠٠٠ إلى ٣٠٠،٠٠٠)، كما يقدّرون أن إيران والهند واليمن يملكون مئات الآلاف من المخطوطات كذلك [٢٣]. وفي الدول العربية، فقد انتشر مؤخرًا في الجامعات والمؤسسات التعليمية الاهتمام بجمع المخطوطات الأصلية وفهرستها، فعلى سبيل المثال تقدّر عدد المخطوطات في جامعة الملك سعود في المملكة العربية السعودية بأكثر من ٣٧ ألف مخطوطة أصلية ومصوّرة. أمّا بشأن المخطوطات المصوّرة، فقد بدأت تظهر جهود أفراد يملكون عشرات أو مئات الألوف من المخطوطات.

الجمع اليدوي للمخطوطات

يقوم بعض الباحثين والمهتمّين بالمخطوطات بالمسح الضوئي لهذه المخطوطات وجمع الصور لها في مواقع على الإنترنت. نذكر على سبيل المثال منها موقع "جامع المخطوطات الإسلامية" [١٥]، وموقع "مركز ودود للفهارس وكتب التحقيق" [٢٤]، وقسم المخطوطات في موقع "ملتقى أهل الحديث" [٢٥]، وقسم

المخطوطات في موقع "مكتبة الألوكة" [٢٦]. إلاّ إن هذه المخطوطات غير مفهرسة، ولا تتيح للباحث البحث في حقول معينة فيها، كما إنَّ الكثير من هذه المواقع لا يملك وصفًا مفصّلًا للمخطوطة، وليست موضوعة في قاعدة بيانات إلكترونية؛ مما يقلل من فائدة هذه المخطوطات في البحث العلمي.

جهود الباحثين في علوم الحاسب الآلي في رقمنة المخطوطات

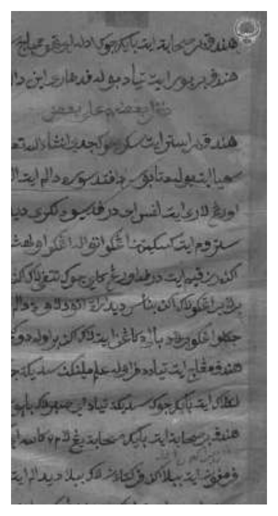
من اهتمامات الباحثين في علوم الحاسب الآلي تطوير البرمجيات المختلفة المعنية بخدمة المخطوطات العربية والإسلامية. إذ من الممكن استخدام قاعدة البيانات في مجال التعرف الآلي على نصوص المخطوطات، وفي تحليل ومعالجة صور صفحات المخطوطات آليًا، والتعرف الآلي على مواقع الكتابة والرسومات والصور والجداول وغيرها وتقسيم صور المخطوطات آليًا إلى مناطق أو أجزاء وتوصيف هذه الأجزاء، كما يقومون بتطوير أساليب البحث الآلي النصي في صور المخطوطات التي تم مسحها ضوئيا، والتعرف على كلمات صور المخطوطات آليًا، بالإضافة إلى تطوير أساليب التعرف الآلي على خط الناسخ، والتأكد منه في المخطوطات العربية والإسلامية. إلاّ إنَّ تطبيق هذه العلوم بحاجة إلى قاعدة بيانات للمخطوطات بمواصفات خاصّة ستتم مناقشتها في الفصل الخامس. لذا قام بعض الباحثين في علوم الحاسب الآلي بمحاولات فردية لبناء هذه القواعد، نذكر في الفقرات القادمة أمثلة عليها، ولكن من الملاحظ أنَّ هذه الجهود عادّة ما تخدم جانبًا ما في علوم الحاسب الآلي مع عجز في خدمة جوانب أخرى، كما إنَّ الكثير منها يحوي عددًا قليلًا جدّا من المخطوطات، بل إنَّ الكثير منها لا يتجاوز جهده المخطوط الواحد فقط.

من الأمثلة على هذه الجهود والمحاولات ما قام به قسم علوم الحاسب الآلي وتقنية المعلومات في جامعة مالاي [٣٠] في ماليزيا إذ جمعوا ١٦٩ مخطوطة من المخطوطات التاريخية والدينية في اللغة الماليزية ولغات الدول المحيطة بها.

بالإضافـة إلى توفـير قاعـدة البيانات لهـذه المخطوطـات على الإنترنـت، إذ قام الباحثون بتوفير البيانات الوصفية المناسبة لكل مخطوطة باستخدام مبادرة ترميز النصوص (TEI) [٢٧] التي ستتم مناقشتها بتفصيل في الفصل الخامس.

أمّـا حسين نفشي وزمـلاؤه في مختبر سينكروميديا التابـع لجامعة كيوبيد في كنـدا [٢٨] فقد جمعوا ١٥ صورة من صـور المخطوطات التاريخية والكتب التراثية الفارسيـة كانـت في خزانة السجـلات التاريخيـة والوثائق والمخطوطـات القديمة لمـيرزا محمـد كاظميني في يزد، إيـران. ويظهر في الصور المجموعـة آثار الرطوبة والطمس والتلاشـي ونزيف الحبر نظرًا لتقـادم الزمان. ومع قلّـة عدد الصور في قاعدة البيانات فإنَّ الباحثين يأملون أن تكون هي النواة الأولى في سلسلة من قواعد البيانـات التـي توفر صور المخطوطـات؛ للمساهمة في علم تحليـل الصور والوثائق الإلكترونية والتعرف على نص الكتابة والكاتب.

كما قام الباحثون رضا مقدّم وزملاؤه في مختبر سينكروميديا وجامعة ماكغيل في كنـدا بتقديم قاعدة بياناتهم المسماة "ابن سينا" [٢٩]، وأدخلوا فيها مخطوط "كشـف التمويهات في شرح التنبيهات" من تأليف سيف الدين أبي الحسن علي بن أبـي علي بن محمد بن سالم الآمـدي (ت٦٣١هـ) وهي تحوي ردًّا وتعليقًا على شرح الـرازي لكتـاب الإشارات لابن سينا. تتكـون قاعدة البيانات هـذه من ٥١ صفحة، قام الباحثـون باستخراج ٢٠٧٢٢ جزءًا من كلمة عربية منها. تمّ التعرف على نص المخطـوط آليًا والتحقق من نتائج التعرّف يدويًا. يظهر رسم توضيحي ٦ أمثلة على جهود الباحثين في علوم الحاسب الآلي المذكورة سابقًا في رقمنة المخطوطات.



(أ) جامعة مالايا ٣٠



(ب) حسين نفشي وزملاؤه ٢٨



(جـ) رضا مقدّم وزملاؤه ٢٩

رسم توضيحي ٦ : صور لبعض المخطوطات من قواعد بيانات رقمية من قبل باحثي في علوم الحاسب الآلي.

جهود المؤسسات والجامعات في رقمنة المخطوطات

تتابعت في الآونة الأخيرة عدد من المؤسسات والجامعات على بناء قواعد بيانات رقمية ضخمة للمخطوطات العربية والإسلامية تمّيزت باحتوائها على مئات أو آلاف المخطوطات، مع توصيف لهذه المخطوطات، وإمكانية البحث في الحقول المختلقة، وتصفّح صور المخطوطات. ورغم أهمّية هذه الخصائص في قواعد البيانات الرقمية، إلاّ إنّ الملاحظ أنَّ هذه الجهود موجهة في الأصل لباحثي العلوم الشرعية مع عدم أخذ باحثي علوم الحاسب الآلي بعين الاعتبار. فممّا يعيب هذه القواعد عدم اكتمال وصف كثير من المخطوطات أو عدم تدقيق ذلك الوصف؛ فقد لا تحتوي على اسم الناسخ مثلًا، أو على حقل خاصّ بنص الصفحة، أو توصيف لأجزاء الصفحات المختلفة. كما إنَّ كثيرًا من هذه المواقع لا تسمح

بتنزيـل المخطوط كامـلاً؛ فكيف بتنزيل مجموعة ضخمة مـن المخطوطات (المئات أو الآلاف) التـي يحتاجهـا باحث علوم الحاسـب الآلي في برمجياته؟!كما إنَّ كثيرًا مـن صور المخطوطات معروضة بدقـة منخفضة، مقارنة بالمعايـير المتَّبعة في مسح المخطوطـات والوثائق التاريخيـة، وعلى الكثير منها علامـات مائية كبيرة الحجم تؤثـر على صورة المخطـوط، ممّا يؤكد الحاجة إلى إعـادة تصميم قواعد البيانات هذه لخدمة الباحثين في العلوم المختلفة على أفضل وجه ممكن.

وهنـا تظهـر مـدى الحاجة في ربـط تواصل بـين المبرمج والباحـث في العلوم الإنسانيـة والشرعيـة، وهو عمل لا يتـمّ إلاّ بوجود تعاون بين مؤسسـات تعنى بعلوم البحث التراثي بالتعاون مع مبرمجين أكفاء يؤمنون بسمو الهدف وثمرته الحقيقية، وهـذه المؤسسـات كثيرة في بلادنـا العربية ولكنهـا تفتقر إلى تنظيم يجمـع آليتها ويجمع جهودها.

نظام المخطوطات لجامعة الملك سعود

نذكر على سبيل المثال في هذا الباب موقع "المخطوطات" التابع لجامعة الملك سعـود في المملكـة العربية السعوديـة [٣١]. وهو نظام قامـت بإنشائه جامعة الملك سعود، وهو متاح للاستخدام من قبل جميع الباحثين في أنحاء العالم. يحوي النظام أكـثر من إحـدى عشرة ألف مخطوطـة، ويستطيع المستخـدم تصفح المخطوطات، وقراءتهـا ومعرفة بعض المعلومـات التفصيلية عنها من مؤلف وعدد صفحات وغير ذلـك. ومن مميزات هـذا النظام أنَّ خدمته تعمل على شبكـة الإنترنت ممّا يسمح للمستخدمين بالدخول من أي مكان، كما إنّ واجهة المستخدم تدعم اللغتين العربية والإنجليزية.

يتميـز نظام المخطوطـات بعرضـ معلومات المخطـوط، بالإضافـة إلى صور المخطـوط كمـا هو مبين في رسـم توضيحي ٧، فتظهـر المعلومات بعضـ التفاصيل عـن المؤلف والكتاب بالإضافة الى معلومـات النـاسخ حين توفرها. كما يسمح الموقع

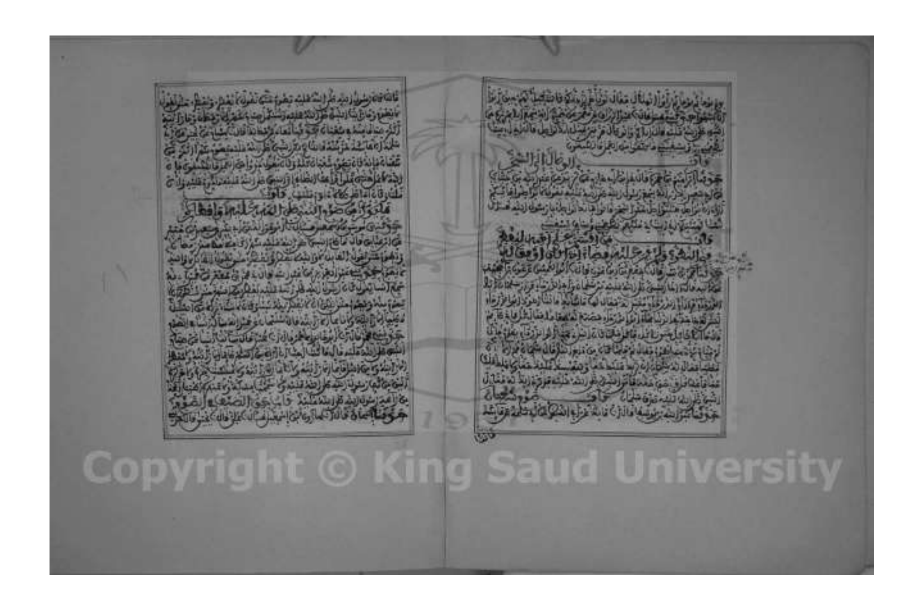
للباحثـين والمهتمين بتفريغ نص المخطوط وإرساله إلى نظام المخطوطات؛ لمراجعته حتـى يتمكن توفير خدمـة البحث في نصّ المخطوط في المستقبـل. لذا يمتلك نظام المخطوطـات في جامعة الملك سعـود عددًا من المقومات اللازمـة للباحثين في علوم الحاسـب الآلي لتطويـر بحوثهـم في خدمة المخطوطـات العربيـة الإسلامية، لكنّه بحاجة إلى عدد من التعديلات والتصحيحات لجعله ملائمًا لهذه الأبحاث.



رسم توضيحي ٧: صورة توضيحية لنظام المحطوطات في جامعة الملك سعود.

وبالرغـم مـن توفر هذه المقومات في نظـام المخطوطات التابـع لجامعة الملك سعـود، إلا إنّ الوضـع الحـالي للنظام لا يسمـح للباحثين باستخـدام المخطوطات العربية والإسلامية. إذ يسمح النظام حاليًا فقط بتصدير كل صفحة من المخطوط علـى انفراد، ولا يسمح بتصدير المخطوط بأكملـه فضلًا عن تصدير مجموعة من

المخطوطـات دفعة واحدة. كما إنَّ الصـور المعروضة ممسوحة بدقة منخفضة (٩٦ نقطـة لكل إنشـ) مقارنة بالمعايير المتّبعة في مسح المخطوطـات والوثائق التاريخية (٤٠٠ نقطـة لـكل إنش) ممّـا يؤثر على قابليـة استخدام هذه الصـور في الأبحاث العلميـة، ناهيك أنَّ القائمين عليها قاموا بإضافـة علامة مائية كبيرة الحجم تؤثر علـى الكتابـة في المخطوط كما هو مبين في رسم توضيحي ٨. وممّا يلاحظ أيضًا أنَّ المعلومـات المقدمة عن المخطوط قليلـة، وفي أحيان كثيرة تنقصهـا بعض البيانات خاصّةً اسم الناسخ.

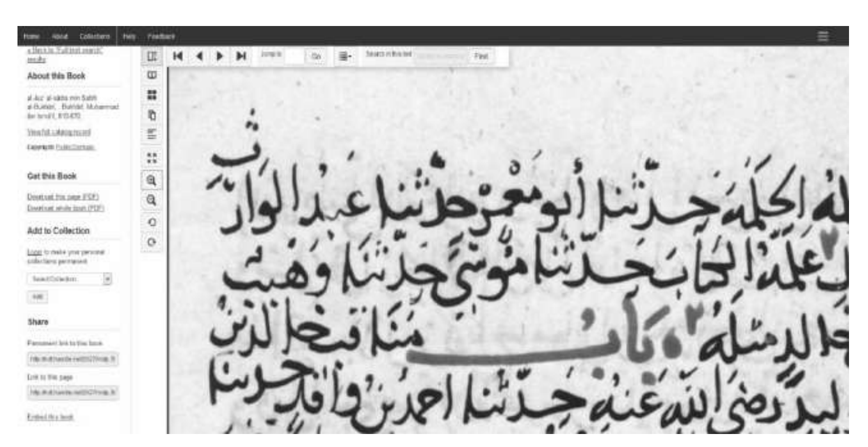


رسم توضيحي ٨: مثال على صور إحدى المخطوطات من قاعدة بيانات جامعة الملك سعود.

مجموعة المخطوطات الإسلامية لجامعة ميتشغان

ومن جهـود الجامعات في توفير قواعد بيانـات إلكترونية للمخطوطات، نذكر "مجموعـة المخطوطـات الإسلامية" في جامعـة ميتشغان الأمريكيـة [٣٢]. تملك الجامعـة ١٧٩٥ مخطوطًـا إسلاميًـا قامـت بعرضـ ٩١٢ مخطوطًـا علـى الشبكة

الإلكترونيــة – إلى تاريخنــا الحــالي – وتوفيرهــا مجانًا للباحثــين والمهتمين، كما يسمح النظام بتحميل صفحة من المخطوط أو المخطوط كاملًا بالإضافة إلى عرض فهرس لكل مخطوط (رســم توضيحي ٩) مع إمكانية التــبرع بإضافة الفهارس أو نص المخطــوط. تتميز مجموعة المخطوطات الإسلاميــة بوضوح صور المخطوطات وجودتهــا، كما إنَّ العلامة المائية موضوعة خارج إطــار صفحة المخطوط؛ فلا تؤثر علــى الأبحاث المعنية بالمخطوط، وممّا يعيبها عــدم شمولية بطاقات المخطوط كما إنَّ الملفــات المحملــة تكون بصيغة (PDF) بدلًا من صيــغ الصور المناسبة للأبحاث العلميــة (JPG, TIF, PNG). ومن الجدير بالذكــر أنَّ موقع مجموعة المخطوطات الإسلاميــة في جامعة ميتشغان على الشبكة الإلكترونية يحوي العديد من المعلومات القيمــة، والقوائــم المفيــدة في اللغــة الإنجليزيــة للمهتمــين بأبحــاث المخطوطات الإسلاميــة والعربيــة، مثــل البحث عن المخطوطــات في الفهارس، وقائمــة بأسماء فهارس المخطوطات، وفهارس المخطوطات الإلكترونية وقواعد البيانات الإلكترونية للمخطوطــات الإسلامية، إضافة إلى العديد من الدراسات المهتمة بالخط العربي، والكثير من المواضيع الأخرى التي لها تعلق بهذا الفن.



رسم توضيحي ٩: مجموعة المخطوطات الإسلامية في جامعة ميتشغان الأمريكية.

يلاحـظ من الأمثلـة السابقة مدى اهتمـام المؤسسات والمكتبـات والجامعات مؤخـرًا في تملّـك المخطوطـات العربية والإسلاميـة ورقمنتها (تحويلهـا إلى صور حاسوبية). وللمزيد من هذه الأمثلة، نحيل القارئ إلى موقع مكتبة جامعة ميتشغان الأمريكيـة [٣٣] فقد ذكـر الموقع قائمة مـن مواقع الإنترنت المحتويـة على قواعد بيانـات للمخطوطات الإسلامية الرقمية، مع ترتيبها علـى حسب الدولة، والجدير بالذكـر أنَّ غالـب هذه الجهود مقـدّم من قبل جامعـات، وأنَّ غالبهـا في دول غير عربية وهي المخطوطات العربية والإسلامية الموجودة في أوروبا، ممّا يدل على حجم المشكلة ومدى تقصير الجامعات والمؤسسات العربية تجاه هذا التراث العظيم.

برنامج مكتبة الملك عبد العزيز العامّة بالرياض

أصدرت مكتبة الملك عبد العزيز العامّة بالرياض [٣٤] برنامجًا رقميًا لعرض المخطوطات التابعة للمكتبة كإصدار أولي تجريبي منذ ما يقارب خمس سنين، على أن يتـمّ تحسـين البرنامج وإصدار نسخة نهائية للجمهـور للاستفادة منه، ولكن لم يتـم اصدار النسخـة النهائية إلى يومنا هـذا للأسف. يحـوي البرنامج ما يقارب ٢٠٠٠ مخطوطـة مختلفـة، ويسمح بالبحـث في اسم المخطوط، أو اسـم المؤلف، أو رقم الحفظ، كما يظهر من رسم توضيحي ١٠. يتميز البرنامج بإمكانية رؤية صور المخطوط مع بعض الخصائص التي تفيد القارئ، مثل تغيير نسبة السطوع، ونسبة التباين، وتكبير الصورة. يفتقد البرنامج إلى توصيف دقيق لبطاقات المخطوطات، فهو يكتفي فقط بتحديد ثلاثة حقول (اسم المخطوط، اسم المؤلف، ورقم الحفظ)، ويعرض باقـي المعلومـات كملف من غـير تنسيق. كمـا إنّ مرور الأعـوام من غير إصدار البرنامج للجمهور أو إعلان تحديثات عليه تقلّل من أهميته في مجال رقمنة المخطوطات العربية والإسلامية.



(أ) شاشة البحث مع الشاشة الرئيسية للبرنامج



(ب) شاشة عرض المخطوط

رسم توضيحي ١٠: برنامج مكتبة الملك عبد العزيز العامّة بالرياض.

مشاريع مستقبلية في رقمنة المخطوطات

ثمّة عدد من المؤسسات التي تذكر مشاريع مستقبلية في رقمنة المخطوطات العربية والإسلامية، نذكر على سبيل المثال منها مؤسسة الفرقان التي أعلنت شروعها في بناء أكبر قاعدة بيانات للتراث الإسلامي على الشبكة العنكبوتية [١٧]، كما قامت جامعة ماكغيل في كندا بتأسيس مشروع "العلوم العقلانية في الإسلام" (Rational Sciences in Islam) [٣٥] والذي يهدف إلى دراسة التراث الفلسفي والعلمي والديني للمسلمين بطريقة شاملة. ويتوقع القائمون على هذا المشروع أن يرفد الباحثين والمهتمين بعشرات الآلاف من المخطوطات الإسلامية بصورة ميسرة وعلمية، ولقد عمدوا إلى تقسيمه لعناصر ثلاثة، منها اثنان خاصّان برقمنة المخطوطات الإسلامية:

- التقاليد العلمية في المجتمعات الإسلامية (STIS): يهدف إلى دراسة تأثير العلوم الإسلامية على العلوم الحديثة في أوروبا، وتأثير الكتابات الغربية ما قبل القرن الثاني عشر الهجري على الباحثين في الدول الإسلامية.

- مبـادرة لبنـاء قاعدة بيانـات في مخطوطات الفلسفـة الإسلامية ما بعد الكلاسيكيـة (PIPDI): يقـوم الباحثـون بإنشاء البنيـة التحتية العلمية المناسبـة لإجـراء دراسـات منهجيـة في النصوص الفلسفيـة الإسلامية التـي يرجع تاريخها إلى مـا بين القرن الخامس إلى القـرن الثالث عشر الهجري. وتشتمل قاعدة البيانات على مخطوطات في المنطق، وفي فلسفة اللغـة، ونظرية المعرفـة، والأخـلاق، والميتافيزيقيا، والفلسفـة الطبيعية وعلم الكونيات، وفلسفة العقل.

- مبـادرة المخطوطات العلميـة الإسلامية (ISMI): يهدف هـذا المشروع إلى توفـير مجموعـة واسعـة مـن المعلومـات حـول العلـوم الدقيقـة في العـالم الإسلامـي قبل العصـر الحديث، من خـلال تقديم هـذه العلوم والمخطوطـات إلى الباحثـين والخبراء بـلا مقابل عن طريـق الإنترنت، وإلى تثقيف الجمهور في جميع أنحاء العالم في العلوم الإسلامية. وينتظر أن تحتـوي قاعـدة البيانات علـى أعمال ما يقـارب ١٧٠٠ مؤلفًـا وعالمًا في مختلـف أرجاء العـلام الإسلامي مـن إسبانيا الإسلاميـة إلى الهند وحدود الصين، منذ بدايـة القرن الأول الهجري حتى القرن الثاني عشر الهجري. وتتراوح موضوعات هذه الأعمال بين علم الفلك، والرياضيات، والفيزياء، والجغرافيا، والميكانيكا، وعدد من التخصصات ذات الصلة.

وبعـد ذكر جهـود ومحاولات الباحثـين والمؤسسات في رقمنـة المخطوطات مع ذكر أمثلة على أبرزها، ومزايا وعيوب هذه المحاولات. ويندرج في هذا الباب كذلك الجهـود المبذولة في محاولة جمـع قواعد بيانات خاصّة بخطـوط العلماء السابقين وناسخي المخطوطات. وهذا ما سيتم تفصيله في الفصل القادم.

جمع نماذج من خطوط العلماء

اهتم العلماء السابقون بتحصيل نماذج من خطوط العلماء والنسّاخ والمشاهير

وجمعها في كتب، وقد دعت الحاجة في زماننا الحالي إلى بناء قاعدة بيانات رقمية تحوي عددًا كبيرًا من الخطوط الموجـودة في الوثائق التاريخية العربية والإسلامية بمعايير محدّدة، مع توصيف كُل هذه الخطوط؛ لاستخدامها من قبل علماء الحاسب الآلي والاستفـادة منها. إنَّ بناء قاعدة بيانات تعد عملية أساسية للأبحاث المتعلقة بتحديـد هوية صاحب الخط والتعرف على الكتابـة في اللغة العربية، إذ من الممكن استخـدام قواعد بيانات الخطوط في الوثائق التاريخية العربية والإسلامية المكتوبة بخط اليد في المجالات البحثية التالية:

١. تحديد هوية كاتب المخطوطات العربية والإسلامية (Writer Identification).
٢. التأكد من هوية الكاتب (Writer Verification).
٣. التعرف على الخطوط المزيّفة وتمييزها، والمساعدة في كشف المخطوطات المزوّرة.
٤. قاعدة بيانات مرجعية للباحثين؛ لاختبار أنظمتهم وأساليبهم المطورة.

ورغـم اعتناء العلماء بخطوط العلمـاء والنسّاخ منذ مئات السنين، وجمع هذه الخطـوط وطباعتهـا في كتـب مفـردة، إلاّ إنّه لا توجـد للآن قاعـدة بيانات متاحة للباحثـين تعنى بخطـوط النسّاخ والكتّـاب في المخطوطات التراثيـة. نذكر في هذا الفصـل أمثلـة على كتب مطبوعة اعتنت بجمع خطوط العلمـاء والنسّاخ مع ميزات هـذه الكتب، ثم نذكر بعض الجهود المبدئية في بنـاء قاعدة بيانات رقمية مختصة بخطوط كتّاب المخطوطات العربية والإسلامية.

كتب مطبوعة اعتنت بجمع خطوط العلماء والنسّاخ

مـن جهود علمائنا السابقين في هذا المجال ما جمعه الحافظ المؤرخ محمد بن علي ابن طولـون الدمشقي "نوادر الإجازات والسماعات" [٣٦] من خطوط علماء مكـة ومصر والشام مـا بين عام ٩٢٠-٩٢٦ هـ. حُقِّق الكتـاب عام ١٤١٩هـ من قبل محمد مطيع الحافـظ. ويحوي الكتاب ست عشرة إجـازة، بالإضافة إلى سماعات

لأحاديـث وتراجم وأسانيد لأمهات الكتب الإسلامية. ويدل هذا الجمع على اهتمام المتقدمين بخطوط علمائهم وتوثيق إجازاتهم وأسانيدهم توثيقا صحيحًا.

ومـن أمثلة جهود علمائنـا المعاصرين قيام خير الديـن الزركلي بتأليف كتابه "الأعلام" [٣٧] إذ حوى ثمانية مجلدات من تراجم المشاهير والأعلام من العلماء في مختلـف الفنون، كالأنبياء، والملـوك والسلاطين، والشعـراء، وغيرهم. كما قام بتوثيق وإضافة صور لعينات من خطوط مئات من الأعلام قديمًا وحديثًا.

كما قام ستيفن ليدر وزملاؤه بجمع وانتخاب السماعات الدمشقية الموجودة في دار الكتـب الظاهرية في مخطوطات عـام ٥٥٠ إلى ٧٥٠ هـ. قام المؤلفون في كتابهم "معجم السماعات الدمشقيـة" [٣٨] بتحليل ١٣٥٠ سماعًا بالإضافة إلى توصيف كل سمـاع بما فيه الشيخ المسمّع، وقـارئ السماع، وكاتبه، وناقلـه، والمستمع، ورقم المخطوط، ورقم السماع، ومكان السماع، وزمان السماع، وغير ذلك.

ومثلهـم جمع عبد الله وجاسم الكنـدري نماذج وأمثلة من خطوط العلماء من القـرن الخامس إلى العاشر الهجري، إذ حـوى كتابهما "خطوط العلماء من القرن الخامسـ إلى العاشـر الهجـري، نماذج وأمثلـة" [٣٩] ثلاثمائـة نموذج لخطوط العلماء من السماعات المختلفة، مع توصيف وتفريغ نص كل سماع. ويظهر أنَّ كتاب الكندريَّـين اعتمـد بكثرة على الاقتباس مـن كتاب "معجم السماعـات الدمشقية" لليدر وزملائه لكثرة النقولات. ويظهر رسم توضيحي ١١ نماذج مقتبسة من الكتب المذكورة آنفًا مع التمثيل ببعض خطوط العلماء.



(أ) كتاب الحافظ ابن طولون الدمشقي ٣٦

(ب) كتاب الزركلي ٣٧

أبو إسحاق جمال الدين إبراهيم بن داود بن ظافر بن ربيعة العسقلاني الدمشقي الشافعي

سمع جميع هذا الجزء وهو الثاني وما قبله وما بعده إلى آخر الكتاب على الشيخ الصالح الأجل المسند نجيب الدين أبي إسحاق إبراهيم بن خليل بن عبد الله الأدمي الدمشقي، بسماعه فيه من ابن الحرستاني بسنده السابق: الأخ الفاضل العالم المحدث علاء الدين أبو القاسم علي بن سيف الدين بلبان بن عبد الله المشرف الناصري، والشيخ الفقيه الإمام العالم الحافظ شمس الدين أبو المحاسن يوسف بن أحمد بن محمود الدمشقي، والشيخ الفقيه صفي الدين أبو عمرو عثمان بن علي بن كاتب الحموي،

(ج) كتاب ستيفن ليدر وزملاؤه ٣٨

(د) كتاب عبد الله وجاسم الكندري ٣٩

رسم توضيحي ١١ : أمثلة من كتب مطبوعة اهتمت بجمع خطوط العلماء.

قواعد بيانات رقمية معنية بخطوط العلماء

اعتنى الباحثون في علوم الحاسب الآلي بالتعرف على الناسخ في المخطوطات الغربية، وكثرت الأبحاث في هذا المجال – نذكر على سبيل المثال منها [٤١ - ٤٣]، ولكنّ اعتنائهم بالمخطوطات العربية والإسلامية يكاد يكون منعدمًا. فعدا عن بحث خاص بصاحب المقال [٤٤]، وبحث آخر عُرِض حديثًا [٤٥]، لا تجد بحثًا مختصًّا - فضلًا عن قاعدة بيانات رقمية مختصّة - في هذا المجال.

قام الباحث في بحثه "التعرف على الكاتب مستقلًّا عن النص في المخطوطات العربية التراثية وتأثير زيادة عدد الكتّاب" " Text Independent Writer Identification of Ancient Arabic Manuscripts and the Effect of Writers Increase " [٤٤] ببناء قاعدة بيانات للوثائق التاريخية العربية والإسلامية المحتوية على عينات مختلفة من تأليف العلماء العرب والمسلمين. تحوي قاعدة البيانات صورًا لمئتي مخطوطة مختلفة، تمّ اختيار خمس صور من كل مخطوطة؛ ليصبح عدد صور قاعدة البيانات ١٠،٠٠٠ صورة لمئتي خطّ ناسخ. واستخدمت قاعدة البيانات المطورة لفحص هوية الكاتب والتحقق منها. وعند استخدام نظام تمييز النمط وبرمجيات التعرف على الكاتب المطوّرة تم تحقيق نسبة نجاح تقديرها: ٩٣,٩٥٪ في التعرف على الكاتب. وتعد هذه النتائج المشجعة دليلًا على قدرة البرمجيات والحاسب الآلي على تمييز الخطوط في المخطوطات والوثائق التراثية. إلّا إنَّ هذه النتائج بحاجة إلى عدّة تجارب معتمدة على قاعدة بيانات تحوي عشرات الآلاف من المخطوطات؛ للتوثق من النتائج، وهو الذي ندعو المؤسسات البحثية التي تُعنى بالتراث وتحرص على إفادة الباحثين إلى تبنّيه وإقامة خطّة لتفعيله على أرض الواقع لما لنتائجه – بإذن الله - من حفظ لتراثنا في صورة علمية وعملية، ولما فيه من تقديم التيسير على باحثين التراث في إنجاز أعمالهم في تحقيق تراث الأمة، ولما فيه من بناء نظام قادر على تقديم خدمة مهمة للباحثين في تمييز خطوط العلماء في المستقبل.

وبعدما قمنا في الفصول السابقة بذكر واقع وحاضر قواعد البيانات الإلكترونية للمخطوطات التراثية، وميزات الجهود الحالية، وأبرز العيوب والعوائق الحالية، نذكر في الفصل القادم مقترحًا لمواصفات قاعدة بيانات مستقبلية للمخطوطات الإسلامية والعربية، وعرضها بالتقنية الرقمية، مع الخيارات والميزّات التي تمكن قاعدة البيانات المقترحة من خدمة باحثي العلوم الإسلامية، وعلوم الحاسب الآلي معًا.

بناء قاعدة بيانات للمخطوطات الإسلامية وتوصيفها

في هذا الفصل نقترح بناء قاعدة بيانات رقمية للمخطوطات الإسلامية والعربية بمواصفات ومعايير عالمية معتمدة؛ تسمح بتطويرها واستخدامها في بحوث الحاسب الآلي، بالإضافة إلى الاستفادة منها من قبل الباحثين والمهتمين بالتراث العربي والإسلامي. تبدأ عملية بناء قاعدة البيانات بمسح صور المخطوطات آليًا وتخزينها على الشبكة العنكبوتية، فيتم مسح هذه الصور باتباع المعايير العالمية في مسح المخطوطات والوثائق التاريخية (بدقة ٤٠٠ نقطة لكل إنش). وبعد القيام بمسح صور المخطوطات آليًا، نقوم بتوصيف رقمي للمخطوط وصوره للاستفادة منه على أفضل وجه. ولتحقيق هذه الغاية، لا بد من تعريف القارئ بمبادرة ترميز وتبادل النصوص الرقمية (Text Encoding Initiative - TEI) [٢٧].

مبادرة ترميز وتبادل النصوص الرقمية

قبل إنشاء مبادرة ترميز النصوص بوصفها نتاج عمل مجلس تعاوني مشترك، لم يجد علماء العلوم الإنسانية معايير موحدة لترميز النصوص الإلكترونية تخدم أهدافهم الأكاديمية. وفي عام ١٩٨٧ م، قام مجموعة من العلماء يمثلون العلوم الإنسانية، واللسانيات، والحوسبة باعتماد مجموعة من المبادئ التوجيهية المعروفة باسم "مبادئ باوكيبسي"، ثمّ تطويرها، وتمّ إعلان مبادرة ترميز النصوص الرقمية [٤٦].

تقوم مبادرة ترميـز النصوص على تحديد مجموعة من العلامات التي يمكن إدراجها في التمثيل الإلكتروني للنصوص، وذلك لتعريف بنية النص ووصفه بطريقة دقيقة، إذ تعتمد برامـج الكمبيوتر على وجود مثل هذه العلامات الواضحة لمعالجة النصوص بطريقة صحيحة، فبدونها يبدو النص الرقمي سلسلة من الأحرف غير المتمايزة، ودون معنى لأجهزة الحاسب.

يعتـبر ترميـز "وصـف المخطـوط" (Manuscript Description) [٤٧] من الوحـدات المتخصصة المندرجة ضمـن نظام "ترميز النصـ الإلكتروني وتبادله" التـي يمكـن استخدامها لتقـديم معلومـات وصفية مفصّلـة عن الوثائـق التراثية والمخطوطـات المكتوبة بخط اليـد. وعلى الرغم من أنَّ عنصـر "وصف المخطوط" صمّـم أصالةً لترميـز المخطوطات في التقاليد الأوروبيـة في العصور الوسطى، فإنَّ برمجيتـه عامّة وشاملة؛ إذ يمكن أن تمتد إلى جميع الوثائق التاريخية بغض النظر عـن لغتها. هـذا الترميز يسمـح للمستخدم بالبحـث عن معلومـات المخطوط من عنـوان، ومواد، وتاريخ ومكان المنشأ. وللتعريـف بعنصر "وصف المخطوط"، نقوم بالتعريف ببعض حقوله:

- msDesc: حقل يتضمن وصفًا أو أكثر للمخطوط.
- msIdentifier: معرّف للمخطوط، يحتوي على المعلومات المطلوبة لتحديد المخطوط.
- head: يحتـوي علـى أي نوع من العناويـن، مثل عنوان القسـم، أو عنوان القائمة، عنوان المخطوط.
- msContents: يتضمن محتويـات المخطوط، كما يصف المحتوى الفكري للمخطوط أو جزء من المخطوط.
- physDesc: يحتـوي حقـل الوصـف المادي علـى الوصف المـادي الكامل للمخطـوط أو جـزء من المخطـوط، ينقسم اختياريًـا باستخدام عناصر أكثر تخصصا.

- history: مجموعـة من العناصر تصف التاريـخ الكامل لمخطوط أو جزء من مخطوط.
- msPart: حقل يحتوي على معلومات حول جزء من المخطوط.
- dimensions: مواصفات أبعاد المخطوطة، أو صفحة معينة من المخطوط، أو جزءًا منها.
- locus: يحدد موضع وإحداثيات موقع داخل جزء من المخطوط.
- material: وصف للمواد التي يتكون منها المخطوط.
- watermark: وصف لأي علامة مائية أو رمز مماثل في المخطوط.
- origDate: يحتـوي على أي شكل من أشـكال التاريخ، وتستخدم لتحديد تاريخ المنشأ للمخطوطة.
- origPlace: يحتـوي علـى أي طريقـة لوصف المـكان، وتستخدم لتحديد مكان الأصل للمخطوطة.

الرموز المطلوبة لتوصيف المخطوطات العربية والإسلامية

نذكـر في الجـداول التالية بعض الحقـول المطلوبة عند توصيـف المخطوطات العربيـة والإسلامية. مـع الإشارة إلى أنَّ بعض هذه الصفـات قد تحتوي على أكثر مـن مُدّخَـل، فعلى سبيل المثال، قد يحتوي المخطوط على عدة عناوين له، وقد يكون للمؤلـف أكـثر من لقب. كما نشـير إلى أنَّ بعض هذه الحقول قـد لا تتوفر في بعض المخطوطـات، أو لا يتـم تعبئتهـا عند توصيفهـا. إلّا إنّ النظام يسمـح بإضافتها أو تعديلها في المستقبل مما يعطي مرونة لقاعدة البيانات المقترحة.

أ. معلومات خاصة بالمخطوط:

عنوان المخطوطة	الموضوع	المستودع	معلومات الكتاب المطبوع
عرض الحاشية	الملخص	حالة المخطوط	صنف الورق
اللغة	عدد اللوحات	أبعاد اللوحات	مقاصد المخطوط
ذكر أول المخطوط	ذكر عبارة الختام	اسم المكتبة	اسم الدولة
اسم المدينة	رقم الحفظ	خط النسخة	رقم الفلم
تاريخ النسخ	تاريخ التأليف	تواريخ أخرى	مصدر معلومات المخطوط
السماعات	الإجازات	التملكات	الوقفيات

ب. معلومات خاصة بالمؤلف / الناسخ / صاحب الحاشية / الخ:

الاسم	الكنية	اللقب	اسم الشهرة
-تاريخ الميلاد	تاريخ الوفاة	مصادر	نوع الخط

ج. معلومات خاصة بصفحة أو منطقة أو جزء من المخطوط:

عدد الأسطر	امتداد السطر	ارتفاع السطر	عدد الكلمات لكل سطر
سمك القلم	عرض الحاشية	إحداثيات محيط المنطقة	نص الكلام في كل سطر
السماعات	الإجازات	التملكات	الوقفيات
الخط المستخدم	الموضوع	رموز أو رسومات	إحداثيات كل سطر

مقترح المهام الخاصّة بأبحاث الحاسب الآلي

وأخـيرًا، نقـوم هنا باقـتراح وتوصيف عدد مـن المهام القادرة علـى استثمار برمجيـات الحاسب الآلي في مساعدة الباحثـين في مجال العلوم الإنسانية والتراث العربي والإسلامي. ولتحقيق هذه المهام، فلا بدّ من سماح قاعدة البيانات المقترحة للمخطوطات الإسلامية للباحثين بتحميل مجموعة من صور المخطوطات ومعلومات هـذه الصـور بأعداد تصل إلى عشـرات الآلاف بسهولة وسرعـة، مما يتطلب أخذ

هذه المتطلبات بعين الاعتبار عند تصميم قاعدة البيانات الرقمية، وهي الميزة التي تفتقدها جميع قواعد البيانات الرقمية للمخطوطات الإسلامية والعربية – للأسف - حاليًا.

كما تتضمن هذه المهام تقسيم المدخلات إلى مجموعات منفصلة للتدريب، والتحقق، والاختبار، وغيرها للمساعدة في تطوير البرمجيات بطريقة علمية دقيقة. مع الإشارة إلى أنَّ عدد الصور أو النصوص المختارة في الفقرات القادمة هي أعداد مقترحة، وهي أعداد تقديرية للتمكن من تحقيق نتائج مرضية على مستوى عالمي في مجال التعرف الآلي، وتعتبر الزيادة على هذه الأعداد المقترحة أمرًا مفضلاً. وتاليًا نحدد الخطوط العريضة لخمسة من المهام المقترحة الخاصة بأبحاث الحاسب الآلي:

أ. **التعرف على الكتابة آليًا** (Automatic Text Recognition)

تعدّ مهمة التعرف الآلي على نصوص المخطوطات من أهم النتائج المتوقعة لعلم الحاسب الآلي في خدمة التراث العربي والإسلامي، وهي في الوقت نفسه من أكثرها صعوبة بلا شك. فإنَّ تحقيق نتائج مرضية في هذا المجال يعدّ شائكًا في صور الوثائق المكتوبة بخط اليد في اللغة العربية عمومًا، وفي المخطوطات التراثية خصوصًا، إذ تحتاج هذه المهمة إلى الكثير من الوقت والجهد لتحضير صور المخطوطات - بالإضافة إلى حاجتها لتوفر نصوص هذه الصور - قبل استطاعة الباحث في الحاسب الآلي البدء في تطوير برمجياته. كما يحتاج إلى تحديد إحداثيات كل سطر في صور صفحات المخطوط (بداية السطر ونهايته وارتفاعه)، وتحديد نص كل صورة سطر على انفراد. والحقيقة أنَّ الجهد الكبير لتجهيز مثل هذه المتطلبات لا بد أن يعتمد على تعاون الباحثين في أنحاء العالم؛ لكي يكتب له النجاح، ولذا تعتمد كثير من الجهود الغربية في مثل هذه الأمور على العمل التطوعي للباحثين عن طريق الإنترنت، ممّا يستلزم بناء الموقع بطريقة تسمح للمشاركة في هذه الجهود، بالإضافة إلى وجود نظام محكَم للتأكد من نتائج هذه الجهود.

بعد تجهيز البيانات اللازمة لمهمة التعرف على الكتابة آليًا، يتم تقسيم الصور إلى أربـع مجموعات منفصلة، نذكرها هنا مـع ذكر قيم تقريبية لعدد المخطوطات في كل مجموعـة (بافتراض وجود ٤٠ ألف صورة للمخطوطات على الأقل في قاعدة البيانـات المقترحـة): مجموعة التدريـب (٥٠٪ من صـور المخطوطات)، مجموعة التحقـق (١٠٪)، مجموعـة الاختبـار (٢٠٪)، والمجموعـة الإضافيـة (٢٠٪). مع ملاحظـة أن يتـم اختيار هـذه الصور من أكبر عـدد ممكن مـن المخطوطات، مع اختـلاف العصور الزمنية، والفنون، ومكان المخطـوط، وأسلوب المخطوط، وغيره، حيـث يسمح التنوع بتطوير برمجيات ذات مرونة قابلـة للتعرف على أكبر عدد من المخطوطات. يستخدم الباحثون عادةً مجموعة التدريب لتطوير برمجياتهم وتعديل معاييرها؛ لتتوافق مع صور المخطوطات المقدّمة لها، ويستخدمون مجموعة التحقق للتأكـد من سلامة نتائجهـم على مخطوطات لم تقـدّم لبرمجياتهم من قبل، ومن ثـمّ يتم اختبار أنظمتهـم على مجموعة الاختبار ومقارنـة النتائج. وتحافظ الجهة المطّـورة لقاعدة البيانات على المجموعة الإضافية من غير نشرها، لاستخدامها في مقارنة نتائج البرمجيات المطوّرة في مسابقات عالمية مستقبلية، إذ تساهم مثل هذه المسابقات على تشجيع البحث العلمي في هذه المجالات وتقدّمه.

ب. البحث عن صور الكلمات (Word Spotting)

يمكـن استخدام صور ونصوص مهمـة "التعرف على الكتابة آليًا" السابقة في تقييم وتطوير مهمة البحث عن صور الكلمات، ولكن مع وضع مجموعة من الكلمات الاختباريـة (١٠٠ – ٥٠٠ كلمة) بالاعتبـار للبحث عنها في هذه الصور. ويفضل أن تكـون هذه الكلمات المختارة من الكلمات المشتهرة عنـد باحثي التراث عند البحث في المخطوطات، ومن الممكن استخدام مهمّة "معالجة اللغة الطبيعية" لمعرفة أشهر هذه الكلمات. وبعد تحديد هذه الكلمات يأتي دور تحديد إحداثيات هذه الكلمات في صـور المخطوطات المختارة لكي يتم تطويـر البرمجيات المختصّة بالبحث عن صور الكلمات.

ج. تحليل صور الوثائق (Document Image Analysis)

يتـم في مهمة تحليـل صور الوثائق اختيـار أكبر تنوع ممكن مـن أشكال صور المخطوطـات، إذ يتمكن الباحثون من الاستفادة من هـذه الصور في تطوير العديد مـن التقنيـات والتطبيقات في مجال تحليل صـور الوثائق، بما في ذلـك على سبيل المثـال لا الحصر: تجزئـة الصفحات والفقـرات والأسطر، والتعـرف على الرموز (مثـل التملكات والأختـام)، والاسترجاع المعتمـد على الشكل، وتمييـز النص عن الرسومـات، والتحليـل المنطقـي للوثائـق، وتحديـد تشابـه الصفحـات، ومعالجة المعلومـات غير النصية. وتحتـاج كلّ من هذه التطبيقـات أولًا إلى تحديد البيانات والمعلومـات المختصـة بهذا التطبيـق. فعلى سبيل المثـال، يستلـزم تطبيق التعرف علـى الرموز تحديد الإحداثيات المئـات أو الآلاف من هذه الرموز الموجودة في صور المخطوطات، مع إعطاء الوصف المناسب لكلٍّ من هذه الرموز.

د. معالجة اللغة الطبيعية (Natural Language Processing)

تعـد هذه المهمة مـن أسهل المهام في جانـب التحضير المبدئـي؛ إذ لا تستلزم صـورًا للمخطوطات، وإنما تحتاج إلى نصوص هذه المخطوطات، والّتي تتوفر بكثرة في المكتبـات الإلكترونية للتراث العربي والإسلامـي. ويفضّل التنوع في اختيار آلاف النصوص لهذه المهمّة من حيث تاريخها، والفن المختار، وما إلى ذلك.

ه. التعـرف على الكاتب والتحقق منه (Writer Identification and Verification)

نحتـاج في مهمـة التعرف علـى الكاتب والتحقـق منه إلى إضافـة ترميز اسم الكاتـب (ورقمـه التسلسلي) لكل صـورة في المخطـوط، مع الاهتمـام بالإكثار من المخطوطات المتنوعة للكاتب نفسه. وهذا يتمّ عن طريق البحث عن النسّاخ المكثرين مـن المخطوطات في قواعـد بيانات المخطوطـات الرقمية. إذ تساعـد المخطوطات المتنوعـة للكاتـب نفسه في التأكد من قـدرة النظام الآلي المطـوّر على التعرف على الكاتـب لا علـى المخطوط أو سنة النسخ. ويفضّل أن يتم اختيار ٢٠ صورة من داخل

المخطـوط (١٠،٠٠٠ مخطوط) مع استثناء صـور البداية والنهاية؛ لاحتوائها عادةً علـى كتابات لغـير الناسخ مثل التملـكات والوقفيات التي تؤثر علـى نتائج البحث، ثـمّ تقسّم الصـور المختارة إلى المجموعات الأربع؛ مجموعـة التدريب (٧ صور لكل مخطوط)، مجموعة التحقق (٣ صـور لكل مخطوط)، مجموعة الاختبار (٥ صور لكل مخطوط)، والمجموعة الإضافية (٥ صور لكل مخطوط).

خاتمة

قـام الباحث في هذا المقـال بتفصيل واقع رقمنة الوثائـق التراثية، مع اقتراح الخطـوات اللازمة في المستقبل للاستفـادة المثلى من البرامج الحاسوبية الحديثة. إذ تمّـت دراسة فهارس المخطوطات الرقمية مع ذكـر أبرز المؤلفات والمواقع المعنية بفهارسـ المخطوطات العربية والإسلامية، وذكر أبرز المزايا والعيوب لهذه المؤلفات والمواقع، وخاصّةً

مكتبة الفرقان الرقمية وفهارس مخطوطات مكتبة المسجد النبوي المتميزتين في هـذا المجال. وكما فصّل المقال جهود الباحثين والمؤسسات في بناء قواعد بيانات المخطوطـات الرقمية، مع إيراد أبرز هذه القواعد وميزاتها وعيوبها، وخاصّةً نظام المخطوطـات التابـع لجامعة الملك سعود الذي نحسبه قـادرًا على أن يصبح مرجعًا معتبرًا في البحث الرقمي في المخطوطات التراثية شريطة زيادة مخطوطاته كثيرًا، وشريطة المراجعة الدقيقة لبطاقات مخطوطاته، وإضافة السمات التقنية الحديثة اللازمـة المبيّنـة في الفصـل الخامس لها أيضًـا. يُلاحظ في جمع الجهـود المبذولة في بنـاء قواعد بيانـات خاصّة بخطوط العلمـاء؛ مدى قلّة هـذه الجهود، والحاجة الماسّـة لتأسيس قاعدة بيانـات رقمية لخدمة بحوث التعـرف الآلي على الكاتب في المخطوطات التراثية.

ونظرًا لعدم قدرة أي من قواعد بيانات الرقمية للمخطوطات التراثية الموجودة حاليًّـا على استيفاء حاجـة الباحثين؛ فقد قـام الباحث بعرض مقـترح لمواصفات قاعـدة بيانـات رقمية مستقبليـة للمخطوطـات الإسلاميـة والعربيـة، مع تحديد الخيـارات والميزّات التي تمكّن قاعـدة البيانات المقترحة من خدمـة باحثي العلوم الإسلامية وعلوم الحاسب الآلي معًا. تحتوي قاعدة البيانات المقترحة للمخطوطات

التراثيــة على مهام محــدّدة مسبقًا تساعد في تطوير البرمجيــات الخاصّة بباحثي الحاسب الآلي: كالتعرف على الكتابة آليًا، والبحث عن صور الكلمات، وتحليل صور الوثائــق، ومعالجة اللغة الطبيعية، والتعرف على الكاتــب والتحقق منه. ونتمنى أن تتواصل الجهود في الاتجاه المقترح لاستكمال بناء قاعدة بيانات رقمية للمخطوطات التراثيــة بالموصفات المطلوبة، حتّى يتسنّــى للباحثين في علوم الحاسب الآلي تطوير البيئــة الحاسوبية والبرمجيات اللازمة لخدمة التراث العربي والإسلامي، والقيام بالواجب المرجو تجاه الأمّة.

شكر وتقدير

يتقـدم معـدّ البحـث بالشكـر لمركز الملـك عبد اللّه بـن عبد العزيـز الدولي لخدمـة اللغة العربية لدعمهم لهذا المقال خاصّــةً وللّغة العربية والدراسات اللغوية الحاسوبيــة عامّةً. ويذكر كذلك بوافر الامتنــان كلّ من ساهم في إثراء هذا البحث مــن أفــكار أو مراجعة أو معلومات قيمة، ونخصّ بالشكــر منهم الباحثين في العلوم الشرعيــة: الشيخ مشهور بن حسن آل سلمان، وسامــي الأسعد، والباحث في العلوم اللغويــة عمــاد السواعير، والباحث في علوم الحاسـب الآلي يوسف العريان. شكرنا أيضًــا موجّه إلى جامعة القصيم لتوفيرها المرافـق الحاسوبية. نسأل اللّه عزّ وجل أن يجعل عملهم خالصًا لوجهه.

المصادر والمراجع

١. الشيــخ مشهـور بن حسن آل سلمــان. الحاسوب وكتب الــتراث الخطية. موقع فضيلة الشيخ مشهور بن حسن آل سلمان. [متصل] ٢٠٠٩. [تاريخ الاقتباس: ١٩ ١١، ٢٠١٤.] http://www.mashhoor.net.

٢. دور تكنولوجيــا المعلومــات في حفظ المخطوطات العربيــة. حافظي زهير. ١٤، القاهرة : البوابة العربية للمكتبات والمعلومات، ٢٠٠٧، Cybrarians Journal.

٣. د. صــلاح الدين المنجد. قواعد فهرسة المخطوطــات العربية. الثاني. بيروت، لبنان : دار الكتاب الجديد، ١٩٧٦.

٤. مجموعــة مساهمــين، تحريــر جيفــري روبر، ترجمــة عبد الستــار الحلوجي. المخطوطــات الإسلامية في العالم - النسخــة العربية. لندن : مؤسسة الفرقان للتراث الإسلامي، ١٩٩٧-٢٠٠٢. صفحة ٣٤٩٦. ٩٧٨-١-٨٧٣٩٩٢-٤٥-٦.

٥. مؤسســة آل البيــت. الفهرســ الشامــل للتراث العربــي الإسلامــي المخطوط (فهارس آل البيت). عمّان : المجمع الملكي لبحوث الحضارة الإسلامية، ٢٠٠٤.

٦. كارل بروكلمــان. تاريــخ الأدب العربــي. [المترجمــون] عبدالحليــم النجــار و رمضان عبدالتواب. الطبعة الخامسة. القاهرة : دار المعارف، ١٩٧٧.

٧. جرجــي زيدان. تاريــخ آداب اللغة العربية. [المحرر] ضيــف شوقي. القاهرة : دار الهلال، ١٩٥٧.

٨. فــؤاد سزكــين. تاريــخ الــتراث العربي. [المحــرر] عرفــة مصطفــى و سعيد

عبدالرحيـم. [المترجمـون] محمود حجازي. الرياض : جامعة الإمام محمد بن سعود الإسلامية، ١٩٩١.

٩. السيـد رزق الطويل. مقدمة في أصول البحـث العلمي وتحقيق التراث. الطبعة الثانية. القاهرة : المكتبة الأزهرية للتراث، ٢٠٠٤.

١٠. معهـد المخطوطـات العربية. فهارس المخطوطـات. الإصـدارات. [متصل] [تاريخ الاقتباس: ٦ ١٢، ٢٠١٤.] http://www.manuscriptsinstitute.org/Esdarat_fahares.aspx.

١١. معهـد المخطوطات العربية. خزانـة المخطوطات. قواعـد البيانات. [متصل] [تاريخ الاقتباس: ٦ ١٢، ٢٠١٤.] http://www.manuscriptsinstitute.org/database_kazana.aspx.

١٢. معالي الشيخ أحمد زكي يماني. مؤسسة الفرقان للتراث الإسلامي. [متصل] [تاريخ الاقتباس: ٠٩ ١٠، ٢٠١٤.] http://www.al-furqan.com.

13. Various Authors، Edited by Geoffrey Roper. World Survey of Islamic Manuscripts. 4th Edition. London : Al-Furqan Islamic Heritage Foundation، 1992-1994. صفحة 2522. 978-1-873992-04-3.

١٤. برنامـج المكتبـة الشاملة. المكتبـة الشاملة. [متصل] برعايـة المكتب التعاوني للدعـوة والإرشاد وتوعيـة الجاليات بحي الروضة. [تاريـخ الاقتباس: ٠٩ ١٠، ٢٠١٤.] http://shamela.ws.

١٥. مشـرف الشهـري. الصفحة الرئيسيـة. موقع جامع المخطوطـات الإسلامية. [متصل] [تاريخ الاقتباس: ٨ ١١، ٢٠١٤.] http://wqf.me.

١٦. فهارس المخطوطات مـن جميع انحاء العالم. ملتقى أهـل التفسير. [متصل] [تاريـخ الاقتباس: ٠٩ ١٠، ٢٠١٤.] http://vb.tafsir.net/tafsir28947/#.VDbETRb-29Z.

17. Al-Furqan Digital Library Portal. Al-Furqan Islamic Heritage Foundation. [متصل] [تاريخ الاقتباس: 09 10، 2014.] http://www.al-furqan.com/world_library/lang/en.

١٨. موقع يوسف زيدان. [متصل] [تاريخ الاقتباس: ٥ ١١، ٢٠١٤.] http://www.ziedan.com.

١٩. مركز الملـك فيصل للبحوث والدراسـات الإسلامية. برنامـج خزانة التراث. [قرص مدمج] الرياض : مركز الملك فيصل، ١٩٩٢.

٢٠. مركـز جمعة الماجد للثقافة والـتراث. برنامج خزانة الماجـد للتراث. [قرص مدمج] الرياض : شركة الدار العربية لتقنية المعلومات.

٢١. الرئاسة العامة لشؤون المسجد الحرام والمسجد النبوي. خدمة البحث بفهارس المخطوطـات. خدمات المكتبـة والباحثين. [متصل] [تاريـخ الاقتباس: ٥ ١١، ٢٠١٤.] http://eservices.wmn.gov.sa/serv_l/L_S004.aspx.

22. Evyn Kropf. Manuscript catalogues online. University of Michigan Library. [متصل] [تاريخ الاقتباس: 28 6، 2014.] http://guides.lib.umich.edu/islamicmsstudies/onlinecatalogues.

٢٣. الشنطـي، عصـام. توزيـع المخطوطـات العربيـة والإسلامية في العـالم. دار التأصيـل. [متصـل] [تاريـخ الاقتبـاس: ٩ ١١، ٢٠١٤.] http://taaseel.com/?PUrl=Bahs&ID=27.

٢٤. موقـع شيخـة المـرّي رحمهـا الله تعـالى. الصفحـة الرئيسيـة. مركـز ودود للمخطوطات. [متصل] [تاريخ الاقتباس: ٩ ١١، ٢٠١٤.] http://wadod.com.

٢٥. قســم المخطوطات. ملتقى أهـل الحديث. [متصل] [تاريـخ الاقتباس: ٩ ١١، ٢٠١٤.] http://www.ahlalhdeeth.com/vb/forumdisplay.php?f=20.

٢٦. سعـد الحميد و خالد الجريسي. قسم المخطوطـات. مكتبة الألوكة. [متصل] [تاريخ الاقتباس: ٩ ١١، ٢٠١٤.] http://www.alukah.net/library/8010.

27. Lou Burnard و Syd Bauman. TEI P5: Guidelines for electronic text encoding and interchange. International Consortium : TEI Consortium.2007،

28. H. Ziaei Nafchi ،R. Farrahi Moghaddam و M. Cheriet. Historical Document Binarization Based on Phase Information of Images [المحرر] .Jong-Il Park و Junmo Kim. Computer Vision - ACCV 2012 Workshops. Berlin، Heidelberg : Springer Berlin Heidelberg.1-12 الصفحات ،2013،

29. IBN SINA: A Database for Research on Processing and Understanding of Arabic Manuscripts Images. R. Farrahi Moghaddam وآخـرون، New York، USA : ACM Press.2010، الصفحات 11-18.

30. Faculty of Computer Science & Information Technology - University of Malaya. Home Page. Digital Library of Malay Manuscripts [متصـل] [تاريخ الاقتباس: 25 7، 2014.] http://mymanuskrip.fsktm.um.edu.my/Greenstone/images/html/mymanuskrip.htm.

٣١. جامعة الملك سعـود. المخطوطات. عمادة التعاملات الإلكترونية و الاتصالات. [متصل] [تاريخ الاقتباس: ٢ ٩، ٢٠١٤.] http://makhtota.ksu.edu.sa.

32. E Kropf. Islamic Manuscripts Collection. University of Michigan Library. [متصل] [تاريخ الاقتباس: 28 6، 2014.] http://guides.lib.umich.edu/content.php?pid=465626&sid=3812450.

33. Evyn Kropf. Online Collections of Digitized Islamic Manuscripts. University of Michigan Library. [متصل] [تاريخ الاقتباس: 15 8، 2014.] http://guides.lib.umich.edu/islamic-msstudies/onlinecollections.

٣٤. مجلس إدارة مكتبة الملك عبدالعزيز العامة. الصفحة الرئيسية. مكتبة الملك عبد العزيز العامّة بالرياض. [متصل] [تاريخ الاقتباس: ٦ ١٢، ٢٠١٤.] http://www.kapl.org.sa.

35. Rational Sciences in Islam. McGill University. [متصل] [تاريخ الاقتباس: 8 11، 2014.] http://islamsci.mcgill.ca/RASI.

٣٦. محمد بن علي ابن طولون الدمشقي. نوادر الإجازات والسماعات. [المحرر] د. محمد مطيع الحافظ. الطبعة الأولى. دمشق : دار الفكر، ١٩٩٨. صفحة ٩٦.

٣٧. خير الدين الزركلي. الأعلام، قاموس تراجم لأشهر الرجال والنساء من العرب والمستعربين والمستشرقين. الطبعة الخامسة عشر. بيروت : دار العلم للملايين، ٢٠٠٢. صفحة ٣٤٩. المجلد ٨.

٣٨. ستيفن ليدر، ياسين السواس و مأمون الصاغرجي. معجم السماعات الدشقية. الطبعة الأولى. دمشق : المعهد الفرنسي للدراسات العربية بدمشق، ٢٠٠٠. صفحة ٥٤٠.

٣٩. عبـدالله الكنـدري و جاسـم الكنـدري. خطوط العلمـاء من القـرن الخامس إلى العاشـر الهجـري، نمـاذج وأمثلة. الطبعـة الأولى. المنامـة : دار البشائر الإسلامية، ٢٠١٤. صفحة ٨٣٩.

40. Using Codebooks of Fragmented Connected-Component Contours in Forensic and Historic Writer Identification. Lambert Schomaker ،Katrin Frankeb و Marius Bulacu. 6 ،U.S.A. : ScienceDirect ،2007 ،Pattern Recognition Letters.28 المجلد ،

41. Automatic Handwriting Identification on Medieval Documents. M. Bulacu و L. Schomaker. Modena، Italy : IEEE، 14 .2007th International Conference on Image Analysis and Processing، ICIAP2007.279-284 الصفحات .

42. An Expert Vision System for Analysis of Hebrew Characters and Authentication of Manuscripts. Laurence Likforman-Sulem ،Henri Maître و Colette Sirat. 2،1991 ، Pattern Recognition.121-137 الصفحات ،24 المجلد ،

43. Hermite and Gabor Transforms for Noise Reduction and Handwriting Classification in Ancient Manuscripts. V. Eglin ،S. Bres و C. Rivero. 2-4 ،New York : Springer101- ، ،122International Journal of Document Analysis and Recognition (IJDAR.9 المجلد ،)

44. Text Independent Writer Identification of Ancient Arabic Manuscripts and the Effect of Writers Increase. Sameh Awaida. Sousse، Tunisia : IEEE2013 ،. ICAI'2013

- The International Conference on Artificial Intelligence.

45. Writer Identification for Historical Arabic Documents. D. Fecker وآخـرون. ،Stockholm، Sweden : IEEE ،August، 2014. Proc. of 22nd International Conference of Pattern Recognition (ICPR.،(

46. Descriptive metadata: Emerging standards. J.R. Ahronheim. 5 ،U.S.A. : Elsevier ،1998 ،Journal of Academic Librarianship.395 صفحة ،24 المجلد ،

47. Text Encoding Initiative. Manuscript Description. P5: Guidelines for Electronic Text Encoding and Interchange. [متصـل]http://www.tei-c.org/release/doc/tei-p5-doc/en/html/MS.html. 16 ،9 2014. [تاريخ الاقتباس: 12 ،11 2014.]

استقصاء تقنيات معالجة اللغات الطبيعية وتطبيقاتها في اللغة العربية

د. أمجد أبو جبارة (*)

amjbara@umich.edu

(*) شركة مايكروسوفت
حصل أمجد أبو جبارة على درجتي الماجستير و الدكتوراة في تخصص معالجة اللغات و استرجاع المعلومات من قسم هندسة و علوم الحاسوب بجامعة ميشيجان الأمريكية في العامين ٢٠٠٩ و ٢٠١٣، كما حصل على درجة البكالوريوس في هندسة الحاسوب من الجامعة الإسلامية في فلسطين في العام ٢٠٠٥. يعمل حالياً كباحث تطبيقي في مجال تطوير محركات البحث في شركة مايكروسوفت الأمريكية . تشمل اهتماماته البحثية حالياً تقنيات استرجاع المعلومات و معالجة اللغات، و تطبيقات تعلم الآلة.

ملخص

يهـدف هذا البحـث إلى التعريف بمعالجة اللغات الطبيعيـة (Natural Language Processing) باعتبارهـا أحـد أهم مجالات علـوم الحاسوب المتفرعة عن مجـال الـذكاء الاصطناعـي (Intelligence Artificial) وباعتبارهـا تمثل نقطة التقاء بينه وبين مبحث اللغويات (Linguistics).

يسلـط البحـث الضـوء على الجهـود البحثيـة التي تستهـدف اللغـة العربية بالتحديـد، مع تبيان مـا للغة العربية مـن خصوصيات يَلزم الباحثـين الانتباهُ لها ومراعاتُها عند نقل التقنيات المستخدمة في اللغات الأخرى إلى اللغة العربية.

يستعرض هذا البحث أهم الموضوعات الفرعية المندرجة تحت مبحث معالجة اللغـات الطبيعية كتقطيع الكلام إلى وحدات، وتصنيـف أقسامه، وإعراب الجمل، وفهـم المعاني، وتوليد الكلام، إلـخ. كما يستعرض الوظائـف الأساسية التي ينبغي للباحثين في هذا المجال الإلمام بها ومعرفة التقنيات والطرق المختلفة لتنفيذها. كما يستعرض البحث -أيضًـا- أهم وأشهر التطبيقات التي تعتمـد بشكل أساسي على تقنيات معالجة اللغات.

يعتمـد هـذا البحث علـى عرض خلاصـات مـا ورد في أهم الكتـب والبحوث المرجعية في المجال، وعقد المقارانات والمفاضلات بين الطرق المختلفة التي اقترحها الباحثون وعرض نتائج تقييم العديد من هذه الطرق.

يتضمـن البحث كذلك استعراضًا مقتضبًـا لأهم الأدوات والمكتبات البرمجية والمجموعـات النصية والبيانية المستخدمة بشـكل متكرر من قبل الباحثين. هذا مع ترك إشارة إلى كل مرجع وكل أداة حتى يستطيع القارئ المهتم اللجوء إلى المصادر للحصول على المزيد من المعلومات التفصيلية في كل موضوع.

١. مقدمة إلى علم معالجة اللغات الطبيعية

هـو أحد علـوم الحاسوب ويعتبر أحد فـروع العلم الناتجة عـن التقاء مبحثي الذكاء الاصطناعـي واللغويات. فعلم الذكاء الاصطناعي يهتم بتطوير خوارزميات وتقنيـات تجعل الحاسوب قـادرًا على تحليل وفهم البيانات التـي يستقبلها، وقادرًا علـى توظيف هذا الفهم لاتخاذ قـرارات ذكية وراشدة في ظروف وسياقات مختلفة –قـد لا يكون تعرضـ لها من قبـل– دون أن يكون قد تمت تغذيتـة بأوامر برمجية محددة تخبره ماذا يفعل في كل ظرف [١].

أمـا علـم اللغويات فهو العلم الذي يهتم بفهم مكونـات اللغة، وأقسام الكلام، والعلاقـات القواعدية والمعنوية بين تلك الأقسام، وأنماط التعبير، وطرق استخدام الألفاظ المختلفة في السياقات المختلفة، إلخ.

وبالتالي فـإن علم معالجة اللغات الطبيعية هو علـم تطبيقي يُعنى باستخدام تقنيـات الـذكاء الاصطناعي – وعلـى رأسها تقنيـة تعلم الآلة – مـن أجل حوسبة المفاهيـم النظريـة الموجـودة في علوم اللغويـات الخاصـة باللغـات الطبيعية التي يستخدمهـا البشر في التخاطـب بحيث يصبح الحاسوب قـادرًا على تحليل وتمثيل وتوليد النصوص المدخلة أو المكتوبة أو المنطوقة بتلك اللغات.

وتدخـل تقنية معالجـة اللغات في كثير مـن التطبيقات الهامـة التي أصبحت تشغـل حيزًا كبيرًا من الحيـاة اليومية لمستخدمي الكمبيوتـر والإنترنت، ومن هذه التطبيقات: محـركات البحث، والترجمة الآلية، والتلخيـص الآلي، والإجابة الآلية عـن الأسئلـة، وأنظمة الإعلانـات الذكية عـبر الإنترنت، وتحليـل المشاعر وقياس الرأي العام، وتصنيـف موضوعات الأخبار، وتنقية البريد الالكتروني من الرسائل الضارة، إلى آخر قائمة تطول من التطبيقات الهامة.

نبذة تاريخية

علـم اللغويات -والذي يستمد مجال معالجة اللغات أسسه النظرية منه- علم قـديم جـدًا يصل عمره إلى مئات السنوات أو ربما أكـثر من ذلك. أما علم معالجة اللغات الطبيعية فظهر بطبيعة الحال بعد وقت قصير من ظهور وانتشار الحواسيب إذ تعود أصوله إلى خمسينيات القرن الماضي عندما وضع العالم الشهير آلن تيورنج Alan Turing معيـارًا يمكن من خلاله الحكم على ذكاء الحواسيب [٢-٣]. يتمثل هـذا المعيار في اختبـار يتنافس فيه الحاسوب مـع إنسان على إجابـة أسئلة يلقيها عليهمـا حكم باستخـدام لغة طبيعية ودون أن يعرف الحكـم مَن مِن المتسابقين هو الـذي يجيب. يعتـبر تيورنج في معياره هـذا أن الحاسوب ذكيٌّ بقـدر كافٍ إذا كان الحكم عاجزًا عن التمييز بين إجابات الإنسان وإجابات الحاسوب.

هـذا المعيار الذي وضعـه تيورنج يستدعي أن يكون الحاسـوب قادرًا على فهم اللغـة الطبيعيـة التي يتلقى فيهـا الأسئلة وأن يكون قادرًا علـى توليد الإجابات بلغة طبيعية كذلك. وقد شكل هذا التحدي أحد أوائل محفزات البحث في مجال معالجة اللغات الطبيعية.

وكان مـن أوائـل التطبيقـات التـي انصـب عليها اهتمـام الباحثـين في فترة الخمسينـات والستينـات من القرن الماضي فكـرة الترجمة الآلية مـن لغة إلى لغة أخـرى وكانت أولى التجارب التي نجحـت نجاحًا مبدئيًا في هذا المجال بحث صدر عـن جامعـة جورج تـاون نجح فيـه الباحثون في ترجمـة قرابة ٧٠ جملـة من اللغة الروسيـة إلى اللغة الإنجليزية [٤-٥]. ومن التطبيقـات الأولى التي لاقت اهتمامًا أيضًـا- أنظمة "إجابة الأسئلـة" Question Answering [٦-٧]، ثم أنظمة تحليل وفهم الحوار [٨-٩]، وأدوات تقطيع الكلام وتحديد أصنافه وإعراب الجمل [١٠-١٣]، ثـم تطبيقات التلخيـص الآلي [١٤]، وأنظمة استرجـاع البيانات [١٥]، ثم تطبيقات تحليل المشاعر التي تزامن ظهورها مع انتشار شبكة الإنترنت [١٦-١٧].

وقـد كان التركيـز في البدايات ينصب بشـكل أساسي علـى تحديد مجموعة مـن القوانـين التي يصوغهـا الباحثون بشـكل يـدوي (Rule-based Approach) ويلقمونهـا للحاسـوب في صـورة برمجية حتـى يتمكن من خـلال تنفيذها من فهم أو توليـد الـكلام، وكان وضع هذه القوانـين يتطلب من البحاثين فهمًـا عميقًا للغة وقواعدها ومعاني كلماتها وكيف يمكن أن تتغير هذه المعاني في السياقات المختلفة.

ولكـن مع ظهور تقنيات تعلم الآلة Machine Learning في أواسط الثمانينات ومـع تطـور سرعـات وذواكر الحواسيـب بدأ الاعتمـاد على وضع قوانـين مصاغة يدويًـا يتراجع بشـكل تدريجي لصالـح الطـرق الإحصائية التي تـترك للحاسوب مهمـة استنبـاط أو تعلم القوانين بشكل آلي من خـلال الاطلاع على كميات ضخمة مـن النصوص والبحث فيها عن خصائص وأنماط متكـررة ذات مدلول إحصائي (كمجيء الصفة بعد الموصوف بشكل متكرر في اللغة العربية مثلًا). وهذا يعني أنه لم يعد من المهم أن يكون لدى مطوري أنظمة معالجة اللغات فهم عميق وكامل لكافة قواعد اللغة ومعانيها، وإنما يلزمهم امتلاك القدرة على تحويل النصوص إلى صيغ إحصائية تتمثل فيها أهم خصائص وأنماط اللغة المتكررة.

مؤخـرًا بـدأت تظهـر بحوث جديـدة تتبنى طـرق هجينة تعتمد علـى كل من القوانـين المصاغـة يدويًـا والأنمـاط الإحصائية التي يتـم تعلمها آليًـا باستخدام تقنيات تعلم الآلة للحصول على نتائج أفضل.

معالجة اللغة العربية

ظهـر الاهتمام بتقنيات معالجة اللغة العربية متأخرًا بقدر ملحوظ عن اللغات الأخرى، وكانت البدايات الأولى في التسعينيات من القرن الماضي، وتناولت الأبحاث وقتها بعض جوانب اللغة كالتحليل الصرفي Morphological Analysis [١٨-٢٠] وتجذير الكلمات stemming [٢١] والتمثيل الحاسوبي لقواعد النحو [٢٢].

وتزايـد الاهتمـام بمعالجـة اللغـة العربيـة بشـكل مطرد مـع دخـول الألفية الثانيــة، لاسيما مــع تزايد الاهتمام بدراسة المنطقة العربيــة ولغتها، وتوفرت أموال وافـرة لدعــم البحوث الخاصة بمعالجة اللغة العربية مــن جهات حكومية وجامعات ومؤسسات بحثية عالمية، كما نُظمت العديد من ورش العمل الخاصة بدراسة تقنيات معالجــة اللغة العربية على هوامش المؤتمرات الكـبرى، وتناولت الأبحاث في أكثرها نقـل التقنيات المستخدمة لمعالجــة اللغات الأخرى وتطبيقها علــى اللغة العربية بعد تكييفهـا وتحويرهـا لتتناسب مع خصوصيــات اللغة العربية، ومن المهــام الرئيسية التــي تناولتها الأبحاث تقطيع الكلام، وتحديد أصناف الكلمات، والإعراب، وتحديد الروابـط المعنوية بين أجزاء الكلام، ونسبــة الضمائر إلى ما تعود عليه. كما تناولت الأبحــاث العديد من التطبيقــات الهامة كالترجمة الآلية بين اللغــة العربية واللغات الأخــرى [٢٣-٢٥]، والتلخيص الآلي [٢٦-٢٧]، وتحليـل المشاعر واكتشاف الآراء [٢٨-٢٩]. هــذا بالإضافــة إلى تناول جوانب بحثية تخص اللغــة العربية كالتشكيل الآلي للكلمات [٣٠-٣٢] وكتمييز ومعالجة اللهجات المختلفة للغة العربية [٣٣-٣٦].

ويميز الباحثون بين ثلاثة أشكال للغة العربية [٣٧]:

أولًا: العربيــة الفصحــى التراثيــة Classical Arabic وهي اللغــة المستخدمة في النصوص الدينية والنصوص التاريخية والكتب القديمة، وتعتبر اللغة العربية القديمــة من أقدم أشـكال اللغة التي حافظت على وجودهــا وانتشارها واستعمالها مــن قبل أعـداد كبيرة من الناس بشكل يومي، ويعود الفضـل في هذا إلى كونها لغة القــرآن الكريم ولغة نصوص السنة النبوية التــي يتداولها ويستخدمها المسلمون – الذين يمثلون أغلبية العرب- بشكل يومي خلال ممارستهم للشعائر الدينية.

ثانيًــا: العربيــة الفصحــى الحديثــة Modern Standard Arabic وهــي اللغة الرسميــة في كل الدول العربية، وتعتبر لغة ثانيــة في بعض الدول غير العربية، وهي اللغــة المستخدمة في الكتابة الرسمية والتعليــم والصحف ووسائل الإعلام. وتمتلك

الفصحـة الحديثة نفس خصائص الفصحى التراثية مـن حيث قواعد اللغة وبناء الجمـل وتصريف الكلمات وأصوات الحروف، ولكنهـا تختلف في كثير من الكلمات والتعابـير، حيث إن ثمة الكثير من الكلمات والتعابـير الموجودة في العربية التراثية تراجـع استخدامهـا في اللغـة الحديثة، ودخلت علـى اللغة كلمـات وتعابير جديدة مقتبسـة من لغات أخرى، أو في إطار الاستجابـة الطبيعية لمتطلبات التطور المعرفي والحضاري.

ثالثًـا: اللهجـات العربيـة العاميـة Dialectal Arabic وتمثـل اللغـات المحلية المستخدمـة في التخاطـب اليومـي بـين الأفـراد في المعامـلات الحياتيـة المختلفة، وهـي تختلف عن بنيـة وقواعد العربية الفصحى كما تختلـف اللهجات العامية عن بعضهـا مـن نطاق جغرافي لآخر ومـن بلد لآخر. وقد اشتهر بـين الباحثين في اللغة العربية تمييز ٦ لهجـات رئيسية في العالم العربي: المصرية، والخليجية، والشامية، والمغاربيـة، والعراقيـة، واليمينية. وتوجد عـدة تقسيمات أخـرى، ولكن يعتبر هذا التقسيـم أكثرهـا شيوعًـا. وكل واحدة من اللهجـات تندرج تحتهـا لهجات فرعية تختلـف مـن منطقة إلى أخـرى، كمـا تختلف بحسـب المستوى الاجتماعـي ومكان المعيشة: المدينة أم القرية أم البادية.

وعليـه فإن اللغة العربية تتمثل فيها ظاهرة يعرفها علماء اللسانيات ويسمونها "ازدواج اللسان" Diglossia وهي تعني أن أهل منطقة معينة يتحدثون لغتين بينهما اختـلاف واضح من حيث القواعد وبنية الجمـل والتراكيب وتصريف الكلمات دون أن يشعروا أنهم يتحدثون لغتين مختلفتين وإنما يتصرفون باعتبارهم يتحدثون لغة واحـدة [٣٨]، مما يوجد تحديـات بحثية لا سيما في ما يتلعـق بمعالجة النصوص المنشورة في وسائل التواصل الاجتماعي التي يكثر فيها الخلط بين العربية الفصحى واللهجة العامية [٣٤].

٢. الوظائف الرئيسية في معالجة اللغات الطبيعية

الفصول التالية من هذا البحث ستتناول المهام والوظائف الرئيسية التي يتكرر الاحتياج إليها في أنظمة معالجة اللغات، مع استعراض مقتضب للطرق المختلفة التي توصل إليها الباحثون للقيام بهذه المهام مع تسليط الضوء على الأبحاث التي استهدفت اللغة العربية بشكل خاص في كل من هذه المهام. كما يتضمن هذا الاستعراض تبيانًا لبعض الوظائف والمهام الإضافية التي تلزم عند معالجة اللغة العربية واللغات الشبيهة بها على وجه الخصوص.

تسوية الكتابة Orthographic Normalization

وتعتبر هذه العملية من أول العمليات التي يتم إجراؤها على النصوص المكتوبة قبل البدء بمعالجتها بهدف التأكد من تنقية النص من الشوائب الكتابية كالرموز الزائدة وعلامات الترقيم غير الهامة لعملية المعالجة مثلًا، وكذلك من أجل التأكد من توحيد الأنماط المختلفة لكتابة الشيء الواحد. وقد تختلف إجراءات التسوية بحسب ما يحتاجه التطبيق. ولكن بشكل عام تتضمن إجراءات التسوية إزالة بعض الرموز الغريبة التي لا فائدة من وجودها في النص، وتوحيد طريقة كتابة علامات الترقيم، وتحويل الحروف الكبيرة Capital Letters في اللغة الانجليزية إلى حروف صغيرة Small Letters إذا كان التطبيق لا يحتاج إلى التفريق بينهما.

وقد بينت بعض البحوث المتعلقة بمعالجة اللغة العربية أن إجراء عمليات التسوية على النصوص العربية له تأثير ملحوظ على جودة وكفاءة عمليات المعالجة اللاحقة للنص [٣٩].

ومن أمثلة عمليات التسوية في اللغة العربية إزالة التطويل من الحروف التي يتم إضافة تطويل لها (عن طريق زر shift + j في الكيبورد مثل: "سيـــــــارة" حيث يظهر فيها طول زائد لحرف الياء) [٤٠]. ومن الأمثلة أيضًا توحيد طريقة كتابة الحروف التي يكثر وقوع الخلط فيها في اللغة العربية كالهمزة (الخلط بين

الوصـل والقطع، والخلـط بين الهمزة على ألف حال كونهـا مفتوحة أو مضمومة — حيـث تكون الهمزة أعلـى الألف- وحال كونها مكسورة - حيـث تكون الهمزة أسفل الألـف -) ، كذلك الخلط بـين الألف المقصورة والياء، وبين الهـاء والتاء المربوطة. فيتـم توحيد كتابة كل هذه الحروف بتحويـل كل حالة يحصل فيها خلط إلى صورة موحـدة في كل النص. ومن أمثلـة عمليات التسوية كذلك إزالة التشكيل (خاصة في حال كونه موجودًا لبعض الحروف أو الكلمات فقط دون أخرى).

النسخ الحرفي Transliteration

ويقصـد بالنسخ الحرفي كتابة الكلمة باستخـدام حروف لغة أخرى غير اللغة الأصلية للكلمة عن طريق تحويل كل حرف إلى حرف أو أكثر يقابله في اللغة الأخرى وغالبًـا ما يكون لفظ هذا المقابـل قريب من لفظ الحرف الأصلي في اللغة الأصلية. مـن أمثلة ذلـك تحويل الحرف "ب" في اللغـة العربيـة إلى "b" أو تحويل الحرف "ش" إلى "sh".

ويقصـد بالنسخ الحرفي في سياق معالجة اللغات كتابة نص لغة ما باستخدام رموز الـ ASCII والتي تقتصر على الحروف الانجليزية الصغيرة والكبيرة والأرقام وعلامات الترقيم وعدد من الرموز الخاصة الأخرى، يتم اللجوء إلى النسخ الحرفي في تقنيـات معالجة اللغات لأسباب عديدة منها وجود أنظمة حاسوب ولغات برمجة لا تدعـم سـوى ترميز ال ASCII، ولا تدعم الترميـز العالمي Unicode الذي يدعم حروف كل اللغات، وحتى في الحالات التي يتوفر فيها الترميز العالمي فمازال العديد من الباحثين الذين يتعاملون مع لغة غير لغتهم يلجأون إلى استعمال النسخ الحرفي لتسهيل التعامل مع اللغة الغريبة عنهم.

واللغـة العربيـة واحدة من اللغـات التي يستعمـل فيها النسخ الحـرفي بشكل متكـرر، وأكـثر قواعد النسخ الحـرفي المستخدمة مـن قبل الباحثـين الأجانب هي طريقـة Buckwalter Transliteration [٤٢]. أو أحـد الطرق المبنية عليها (مثال

Habash-Soudi-Buckwalter [٤٢]). ووفـق طريقـة تحويـل Buckwalter يكون لـكل حـرف عربي ما يقابله (قد يكون حرف أو أكـثر أو حرف و رمز) في ترميز ال ASCII، فمثـلاً كلمة "سماء" تنسخ حرفيًا إلى " samaA`"، وكلمة "كان" تنسخ إلى "kaAna"، وهكـذا. ويكـون التحويل لكل كلمة تحويلًا فريدًا بحيث يمكن عكس عملية التحويل لاستعادة الكلمة الأصلية كما هي.

ونحيـل المطوريـن المهتمين بتطوير وبرمجـة أدوات معالجة اللغـة العربية إلى البرنامـج المشار إليه في [٤٣] والذي يقوم بتحويل النص العربي المرمز باستخدام الـ Unicode إلى ترميز ASCII وفق طريقة Buckwalter وبالعكس. بالإضافة إلى هذا فإن الواجهة البرمجية (API) المرفقة بـ"المجموعة النصية العربية القرآنية" تتضمـن برنامجًا يقـوم بنفس المهمة باستخدام طريقـة مشتقة من طريقة Buck-walter بعد إدخال تغييرات طفيفة عليها [٤٤].

تقطيع الكلام Tokenization

وتسمى هـذه العملية أيضًا بالتحليل اللفظـي Lexical Analysis، ويقصد به تقطيـع النص إلى وحـدات Tokens تتكـون كل وحدة منها من أحـرف أو أرقام أو رمـوز متصلة كالكلمـات أو الأعداد أو علامـات الترقيم، مع تحديـد موضع بداية ونهايـة كل وحدة. وتختلـف درجات التقطيع التي يتم إجراؤهـا على النص بحسب مـا يتطلبه التطبيـق، ومنها "التقطيع البسيط" وغالبًا مـا تستخدم فيه التعبيرات النمطيـة Regular Expressions لتقطيـع النص إلى الوحـدات التي تفصل بينها المسافـة Whitespace أو نهايـة السطـر، كذلك يتـم فصل علامـات الترقيم عن الكلمـات (كالفاصلة المتصلة بالكلمة)، بحيث تصبح الكلمة وحدة وعلامة الترقيم وحـدة أخرى منفصلـة عنها، وكذلك فصل الأرقام عن الكلمـات إذا كانت متصلة بهـا، وكذلك فصل بعض أدوات الربط التي قد تتصل بالكلمات أحيانًا كما في اللغة العربيـة مثـلاً حـين يتصل حرف العطـف "و" بالكلمـة التي تليه كمـا في كثير من

النصوص العربية. وقد تشمل عمليات التقطيع البسيط إزالة بعض الرموز الغريبة أو استبدالها كلها برمز واحد بهدف تقليل درجة التشويش noise في النص.

وقد تتطلب بعض التطبيقات عملية "تقطيع متقدم" تذهب إلى أبعد من مجرد تقطيع النص إلى كلمات تفصلها مسافة Whitespace [٤٥-٤٦]، وفيها يتم تقطيع الكلمة نفسها إلى وحدات أصغر إذا كانت الكلمة ناتجة عن تركيب عدة مكونات كدخول "ال" التعريف على الكلمة أو الضمائر المتصلة كما في اللغة العربية. وفي هذه الحالة تتطلب علمية التحليل إجراء تحليل صرفي Morphological Analysis للنص للتعرف على المكونات التي تتركب منها الكلمات حتى يكون من الممكن تقطيعها على هذا النحو.

الشكل ١ يوضح مثالاً لجملة باللغة العربية ونتيجة تقطيعها تقطيعًا بسيطًا وتقطيعًا متقدمًا، وتوجد بين هذين المستويين من التقطيع مستويات عدة يتم اللجوء إلى ما يناسب التطبيق من بينها [٣٧].

<table>
<tr><td>الجملة</td><td>سيلقي المدير كلمة، وستتناول كلمته شرحًا للخطة الجديدة.</td></tr>
<tr><td>تقطيع بسيط</td><td>سيلقي | المدير | كلمة | ، | و | ستتناول | كلمته | شرحًا | للخطة | الجديدة | .</td></tr>
<tr><td>تقطيع متقدم</td><td>س | يلقي | ال | مدير | كلم | ة | ، | و | س | تتناول | كلم | ة | ـه | شرح | ل | ال | خط |
ة | ال | جديد | ة | .</td></tr>
</table>

شكل١ : مثال يوضح كيفية تطبيق عمليات التقطيع البسيط والمتقدم على نص عربي.

جدير بالذكر أن عملية تقطيع النص –حتى في أبسط أشكالها- تعتبر عملية صعبة ومعقدة للغاية في اللغات التي لا تستخدم فواصل محددة بين الكلمات ومن أمثلة هذه اللغات اللغة الصينية واليونانية التراثية واللغة التايلاندية وغيرها.

ومن الأدوات التي يمكن أن يستعملها الباحثون والمطورون لتقطيع النص العربي أداة TOKAN [٤٧]، وهي تحتاج إلى إجراء تحليل صرفي للنص حتى

تتمكـن من تقطيعه، وهو مـا توفره أداة MADA وهي متاحـة مع TOKAN كباقة واحدة للتحميل من [٤٨]، وتحتوي الباقة -بالإضافة إلى المحلل الصرفي والمقطع- على أدوات أخرى، منها أداة لتسوية النص Orthographic Normalization، وأداة لتحويل النصوص العربية إلى ترميز ASCII وفق طريقة Buckwalter كما شرحنا في الفصل السابق، وغير ذلك.

ومن الأدوات المتاحة أيضًا AMIRA-TOK وهي إحدى الأدوات التي تتضمنها باقة AMIRA [٤٩] التي تضم أدوات لتنفيذ العديد من المهام الأساسية في معالجة اللغة العربية، وهي متاحة للتحميل من [٥٠]، و تصل دقة التقطيع في AMIRA إلى F1 = ٩٩,٢٪ مـع العلم بأن التقطيع الذي يجريه البرنامج لا يتعامل مع كل حالات التصريف وبالتحديد ما كان داخليا كالـ Inflectional Morphology.

التحليل الصرفي Morphological Analysis

يقصـد بالصرف Morphology تحويل الأصل أو الجذر من الكلام إلى أبنية وأشكال مختلفة تحمل دلالات معنوية مرتبطة بمعنى الجذر، ومن أشكال الصرف: الجمـع والتثنية والتأنيـث واسم الفاعل واسـم المفعول والصيـغ الزمنية للأفعال: مضـارع وماضي وأمر، وصيغ المبالغـة، إلخ [٥١]. ومثال على ذلك في اللغة العربية الجـذر "ذهب" الذي يمكن أن يأتـي في التصريفات التالية: "يذهب" و"يذهبون" و"يذهبان" و"تذهبان" و"الذاهب" و"الذاهبون" و"الذاهبات" ... إلخ.

وتهدف عملية التحليل الصرفي للكلمات إلى دراسة بنية الكلمة بغرض التعرف علـى القسم الصـرفي لها، كتحديد هل هي جمع أم مفـرد، صيغة تذكير أم تأنيث، صيغـة ماضٍ أم مضارع أم أمر للأفعال ... إلخ، كما تهدف إلى تحديد جذر الكلمة وتحديد الزوائد التي أدخلت على الجذر.

ويميز الدارسون لعلم التحليل الصرفي بين نوعين من التحليل [٣٧]:

أولا: تحليـل معتمد على النوع أو الشكل Form-based Morphology : ويقصد بـه تحديد مجموع "المورفيمات" Morhpemes المكونة للكلمة وأنواعها. و"المورفيم" هـو أقل لفظة في اللغة تفيد معنـىً متضمنًا فيها (قد تكون جزءً من كلمة – كضمير متصل مثلًا). والمورفيمات التي تتكون منها الكلمات تشمل الجذع Stem مثل "كتب" في كلمـة "تكتبـين"، والاضافـات السابقـة prefixes مثل "التاء" التـي تدخل على الجذع للدلالة على كون الفعل مضارع للأنثى الغائبة أو المذكر المخاطب، والإضافات اللاحقـة suffixes مثل "ين" فتصبح الكلمة "تكتبـين" للدلالة على توجيه خطاب للأنثـى في صيغـة الفعـل المضـارع. والإضافـات الداخليـة infixes كدخـول الألف علـى الفعـل فتحوله إلى صيغة اسـم الفاعل كما في "كتـب" - "كاتب"، ويعتبر من المورفيمـات أيضًا الزوائد التي تحمل خصائص كلمـة منفصلة ولكنها تأتي متصلة مع الكلمة في الكتابة مثل "أل" التعريف والضمائر المتصلة مثل "هم" في "كتابهم".

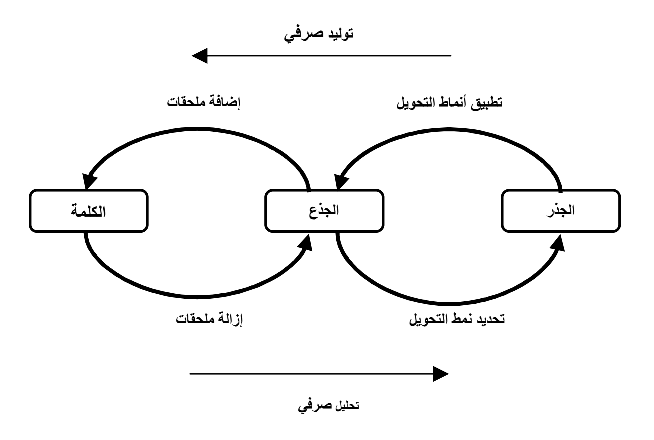
ثانيًـا: تحليل معتمد على الوظيفـة Function-based Morphology: ويقصد بهـا تحديد الدور الإعرابـي والدلالة المعنوية التي تنتج عن دخـول كل مورفيم على الكلمة، وتنحصر وظائف المورفيمات المختلفة في ثلاثة أنواع رئيسية:

أولا: تصريـف اشتقاقـي Derivational Morphology: ويقصد به أن دخول المورفيمـات علـى الكلمة يصنع منها كلمـة جديدة، فمثلًا "جامـع" أصلها "جمع" ودخـول مورفيـم زائد عليهـا جعلها تعنى مـكان الصلاة في المعنى الـدارج للكلمة، و"جامعـة" كذلـك أصلها "جمـع" ودخل عليهـا مورفيمان فأصبحـت تعني مكان للتدريس في المعنى الدارج للكلمة.

ثانيًـا: تصريـف إعرابـي Inflectional Morphology: ويقصـد بـه دخـول المرفيمات على الكلمة بغرض تحديد الزمن، أو الجنس، أو حالة الخطاب، أو العدد كما أوضحنا في الأمثلة السابقة.

ثالثًـا: تصريـف إلحاقــي Cliticization Morphology: ويقصد به أن يلحق بالكلمة مورفيم يحمـل الخصائص المعنوية والإعرابية للكلمات المنفصلة. وضربنا آنفًا مثال دخول أل التعريف على أول الكلمة أو أحد الضمائر المتصلة على آخرها.

الشكل ٢ يوضـح نظام الاشتقاق والتحويل بين الكلمــة وجذعها وجذرها ودور عمليات التصريف في هذا الاشتقاق [٥٢].



شكل ٢ : نظام الاشتقاق الصرفي والعلاقة بين الكلمة وجذعها وجذرها

ولعلـم الصــرف في اللغات المختلفــة قواعد وقوانين تضبـط عمليات الصرف وتحدد أشكال وأنــواع الصرف التي تصلح لكل سياق وطريقة تنفيذ هذا الصـرف. وبقــدر ما تكـون هذه القواعد قليلة ومحددة تكون حوسبــة التحليل الصرفي أسهل. وتعتبر اللغة الانجليزية مثلًا من اللغات التي لها نظام تصريف سهل. في حين تعتبر اللغــة العربية من اللغــات الغنية من الناحية الصرفية بمعنــى أن هناك الكثير من الأشــكال والأبنية الصرفيــة التي يمكن تكوينها من الجــذر الواحد مقارنة بمعظم اللغــات الأخـرى، وهنـاك الكثير من القواعــد والقوانين التي ينبغــي مراعاتها عند تصريــف الكلمات. ولذلك يعتبر التحليل الصرفي من أبــرز وأهم المهام التي يُعنى

بهـا الباحثون في مجـال معالجة اللغة العربيـة [٣٧] ولذلك تجـد أن بحوثًا كثيرة قـد تناولت هذه الجزئيـة (أمثلة:[٢٠، ٥٣-٥٤]) واقترحت طـرق متعددة لإجراء التحليـل الصـرفي للنصوص العربيـة. ويمكـن للقارئ المهتـم أن يرجـع إلى أحد الاستقصـاءات المنشـورة في هذا الموضوع والتـي حصرت الطـرق المختلفة وقارنت بينهـا (مثـال: [٥٢]). حيـث اقـترح المؤلفون تصنيـف الطرق المختلفـة لحوسبة عمليات التحليل الصرفي إلى أربعة أقسام:

أولا: البحـث في جدول شامـل Table Lookup Approach: و فيها يتم تخزين كل الكلمات العربية الممكنة في جدول كبير يحتوي أيضًا تفكيك كل كلمة إلى مكوناتها من المرفيمات. هذه الطريقة تحتاج إلى جهد بشري كبير لبناء هذا الجدول.

ثانيًـا: الطريقة اللغويـة Linguistic Approach: وفيها يقـوم خبراء في اللغة وقواعد الصرف فيهـا بصياغة مجموعة من القوانين التي تحكم عمليات الصرف لأكبر عدد ممكن من الكلمات. ويتم تغذية الحاسوب بهذه القوانين حتى يستخدمها لتحليل الكلمات [٥٥].

ثالثًـا: طريقة التوافيـق Combinatorial Approach: وفي هـذه الطريقة يتم تجريـب كل المجموعات التي من الممكن تشكيلهـا من حروف الكلمة، ويتم بعد ذلك مقارنـة كل مجموعـة من تلك المجموعـات بقائمة شاملة بالجـذور العربية roots، وقوائم للملحقـات suffixes [٥٦-٥٧]. وهذه الطريقـة مكلفة من حيث احتياجها لطاقة حوسبة أعلى حتى تتمكن من اختبار كل المجموعات الممكنة في وقت معقول.

رابعًا: طريقة الأنماط Pattern-based Approach: وتعتمد هذه الطريقة على تحديـد عدد من الأنمـاط المتكررة في الكلمات العربية مـن حيث التركيب الصرفي وتطبيق هذه الأنماط على الكلمات من أجل تحليلها [٥٨].

ونحيـل القارئ المهتـم إلى المصادر المشار إليها في هـذا الفصل للتعرف أكثر على الطرق المستخدمة في التحليل الصرفي للكلمات العربية.

ومـن أدوات التحليـل الصرفي المتاحـة للباحثين والمطوريـن باقتا MADA و AMIRA اللتـان تمـت الإشـارة إليهما في الفصـل السابق، ومنهـا أيضًا BAMA (Morphological Analyzer Buckwalter Arabic) [41,59] و SAMA (-Standard Arabic Morphological Analyzer) [٦٠] وهـو مبنـي على BAMA إلا إنه لا يتوفر مجانًا حاليًا.

ومـن الأدوات أيضًـا ElixirFM ويتوفر له عرض لتجريبـه من خلال واجهة يمكن تشغيلها عبر المتصفح [٦١]، كما أنه متاح للتحميل المجاني [٦٢].

التجذيع Stemming

هـي عملية يتـم إجراؤها علـى الكلمة وتهـدف إلى تحويل الكلمـة إلى صورة أبسـط بحيث تكـون أقرب ما يمكن إلى جذع أو جذر الكلمـة وذلك من خلال إزالة الزوائـد الصرفيـة التي تدخل على أصـل الكلمة. فمثلًا جذع كلمـة computer في اللغة الإنجليزية هو comput، وكذلك جذع كلمة computing هو comput، ومثلهم أيضًـا كلمات computed و computers و computation كل هذه الكلمات لها جذع واحد هـو comput. ونلاحظ هنا أن comput في حد ذاتها ليست كلمة صحيحة في اللغة الانجليزية وليست هي الجذر الذي اشتقت منه الكلمات الأخرى ولكنها قريبة جدًا من الجذر.

وتقـوم عملية التجذيع هذه على افتراضـ أن الكلمات المشتقة من نفس الأصل تحمـل نفسـ المعنى أو معاني متقاربة جدًا في الغالبيـة العظمى من الحالات، وعليه فإن الهدف الأساسي من عملية التجذيع هو تحويل كل كلمة في النص إلى أقرب ما يمكن من أصل هذه الكلمة بحيث لا يعود هنـاك اختلاف بين computer و computers و computing، وعند مقارنة النصوص لاحتساب مدى تشابه نص مع نص آخر يتم التعامل مع كل هذه الكلمات على أنها كلمة واحدة.

ويتـم اللجوء إلى عمليـة التجذيع في كثير من تطبيقـات معالجة اللغات، ومن أمثلـة ذلك عمليـات الفهرسة indexing التي تقوم بها أنظمـة استرجاع المعلومات Information Retrieval كمحـركات البحث وما شابههـا، فأنت عندما تبحث عن كلمـة Computers عـبر أحد محركات البحث، فلن يكون من الخطأ أن يعرض لك محـرك البحث نتائج فيها كلمة Computer أو Computing [٦٣-٦٤]. ومثل ذلك أيضًا التطبيقات التي تحتاج إلى احتساب مقدار التشابه بين النصوص، كتطبيقات تصنيف النصوص Text Classification، وتطبيقات تجميـع النصوص المتشابهة Text Clustering، وكذلـك تطبيقات كشف النسخ في المنشورات Plagiarism Detection.

وتعتمد أغلب خوارزميات تجذيع الكلمات على مجموعة من القواعد والقوانين المعـدة بشكل يدوي، وتختلف هذه القواعد من لغـة إلى لغة، ولكن الرابط المشترك بينها كلها أنها تحاول أن تزيل الزوائد الصرفية التي تدخل على الكلمة. ويعتبر هذه النوع مـن الخوارزميات الأشهر والأكـثر استخدامًا بسبب سهولـة تطبيقه وسرعة تشغيله على النصوص وعدم حاجته إلى قوائم كلمات أو قواميس. وتعتبر خوارزمية Porter التـي اقترحها Matrin Porter (عام ١٩٨٠) أشهر هذه الخوارزميات للغة الانجليزية ومازالت مستخدمة على نطاق واسع حتى اليوم [٦٥].

وهنـاك مجموعة أخرى –ولكنها أقل شيوعًا– من خوارزميات التجذيع تعتمد علـى رصد خصائص إحصائية للكلمات، وذلك من خلال إيجاد الكلمات المتشابهة التـي تظهر معًا بشـكل متكرر في النصوص المختلفة ثم اختيـار الكلمة الأقصر من بـين كل هـذه الكلمات لتكون هي الجـذع الذي ترد إليه كل الكلمـات الأخرى [٦٦-٦٧]. ولكـن هذه الطريقـة تحتاج إلى توفر كميات كبيرة جـدًا من النصوص تكفي لاحتساب علاقات إحصائية ذات معنـى بين الكلمات. والميزة الأساسية لهذا النوع مـن الخوارزميات هو قدرته على تجذيع التصريفات الغير قياسية مثل foot و feet

وهـو ما تفشل الخوارزميات المعتمدة على القواعد المصاغة يدويًا في التعامل معه. الميزة الاخرى لهذا النوع من الخوارزميات هو أن الخوارزمية نفسها يمكن تطبيقها علـى أكثر من لغة. في حين أن القواعد المصاغة يدويًا تختلف بحسب اللغة وتحتاج إلى خبير لغوي حتى يتمكن من تحديد تلك القواعد.

اللغـة العربية تعتبر من اللغـات التي يصعب تطبيق التجذيع فيها لعدة أسباب أهمهـا الـثراء الصرفي لهـا وكون الزوائـد الصرفية تدخـل على الكلمـة في أولها ووسطهـا وآخرها، مما يجعل صياغـة قواعد للتجذير عملية صعبـة ومعقدة جدًا، هذا بالإضافة إلى الجدل بين الباحثين حول درجة التجذيع اللازمة في اللغة العربية بحيـث تحقق مـا تحتاجته تطبيقات استرجاع المعلومـات أو تصنيف النصوص بين مـن يـرى أن الاقتراب من صـورة الجذع يكفي وبـين من يرى ضـرورة إيجاد جذر الكلمة الثلاثي أو الرباعي [٦٨].

وقـد اقترحت عدة طرق لتجذيـع الكلمات العربية منها مـا يعتمد على قوائم معـدة يدويًا تحتوي كل كلمة وجذعها (أو جذرها) [٦٨-٦٩]، ومنها ما يعتمد على صياغـة قواعـد شبيهة بما تقـوم به خوارزميـة porter في اللغـة الانجليزية (مثال [٥٨])، ومنها كذلك ما يعتمد على الطرق الإحصائية [٧٠-٧١].

ومـن أدوات التجذيـع المتاحـة للاستعمال من قبـل الباحثـين والمطورين أداة Khoja Stemmer [٧٢] التـي تستخـدم مجموعـة مـن القوانـين المصاغة يدويًا لحذف الزوائد، ومجموعة من الأنماط والقوائم لاستخراج الجذع (أو الجذر).

تأصيل الكلمة Lemmatization

وهي عملية شبيهة جدًا مـن حيث الهدف بعملية التجذيع Stemming ولكنها تختلـف عنها من عدة نواحي أهمهـا أن الأصل الناتج عن العملية هو كلمة صحيحة في اللغـة وتكـون في الغالب هـي أبسط شكل للكلمـة، ففي المثال الـوارد في الفصل

السابق، فإن أصل كلمة computers و compute (قارنه بالجذع comput). الفارق الآخر هو أنها تقوم بتحويل الكلمة إلى كلمة أخرى إذا استدعى الأمر ذلك بشكل لا تستطيع أغلب خوارزميات التجذيع القيام به. مثال ذلك تحويل is وare و am إلى be وكذلك better إلى good وهكذا.

وتعتمد أغلب خوارزميات تأصيل الكلمات على استخدام قوائم بالكلمات المختلفة وأصولها مع دمج ذلك مع بعض القوانين البسيطة لجعل حجم تلك القوائم أصغر.

تصنيف أقسام الكلام Part of speech tagging

وهي واحدة من العمليات الأساسية اللازمة للكثير من تطبيقات معالجة اللغات، ويتم فيها تصنيف كل كلمة في النص بحسب حالتها الصرفية وبحسب سياقها الإعرابي، كتصنيف الكلمة كفعلٍ ماضٍ أو اسم بصيغة الجمع أو أداة عطف ... إلخ.

والصعوبة الأساسية في هذه العملية تكمن في ضرورة أخذ السياق الذي ترد فيه الكلمة في عين الاعتبار وذلك بسبب الغموض الذي يكتنف بعض الكلمات إذا ما عوملت منفصلةً عن سياقها. مثال لذلك في اللغة العربية كلمة "ذهب" ففي بعض السياقات هي اسم معدن ثمين، وفي سياقات أخرى هي فعلٌ ماضٍ للمفرد الغائب.

يعرض شكل ٣ مثالين أحدهما باللغة الإنجليزية والآخر باللغة العربية يوضحان نتيجة تطبيق عملية تصنيف أقسام الكلام على النص. الرموز الظاهرة في المثالين مثل VBZ و NN و PRP$ وغيرها تسمى Tags، وكل منها يرمز إلى أحد أقسام الكلام Part-of-speech ، فمثلًا VBZ تشير إلى الفعل المضارع للمفرد في اللغة الإنجليزية.

My son also likes eating meat ./. My/PRP$ son/NN also/RB likes/VBZ eating/VBG meat/NN تعجبني الأفكار الإبداعية ويسعدني أن أساهم في إنجاحها تعجبني/VBP الافكار/DTNN الابداعية/DTNN و/CC يسعدني/VBP ان/IN اساهم/VBP في/IN انجاحها/PUNC ./NNP

شكل ٣: نتيجة تطبيق عملية تصنيف أقسام الكلام على جملة إنجليزية وأخرى عربية

ويتـم تعريف الأقسـام Parts-of-speech التي يتم تصنيف الكلمات إليها من قبل علماء اللغويات، الذين يعرفون مجموعة من الأقسام تسمى TagSet. ولكل لغة مـن اللغات أصناف مختلفة من الكلام تختلف بحسـب الثراء الصرفي والقواعدي لتلـك اللغة. كمـا أن اللغة الواحدة قد يلجأ العلماء فيهـا إلى تعريف أكثر من Tag-Set تختلـف في درجـة تفصيلها. فمثـلًا أبسط TagSet في اللغـة العربية ممكن أن تحتـوي ثلاثة أقسام فقط: اسم وفعل وحـرف. في حين أنه من الناحية النظرية ممكن تعريف عدد قد يزيد على ٣٣٠،٠٠٠ قسم!

ومـن أشهر التقسيمـات المستعملـة في اللغـة العربيـة Buckwalter TagSet بصورتيهـا المخصصة للكلمـات المقطعة Tokenized (وهي الأكـثر شيوعًا وتضم قرابـة ٥٠٠ قسـم) وغـير المقطعـة Untokenized وتصـل أقسامهـا إلى عشرات الآلاف. وفي حالـة تقسيـم الكلام Tokenization فإن الكلمـة الواحدة تنقسم إلى مكوناتها الصرفية وكل من تلك المكونات يكون تابعًا لقسم مختلف.

هـذا بالإضافة إلى العديد من التقسيمات الأخرى التي حاول بعضها اختصار Buckwalter TagSet بهـدف تقليـل عـدد التقسيمـات وتسهيل عمليـة التصنيف [٧٣-٧٦].

وتستخـدم معظـم خوارزميـات تصنيـف أقسام الـكلام حاليًا تقنيـات تعلم الآلــة، وذلك مــن خلال إعـداد كمية كبيرة مــن النصوص والاستعانــة بخبراء في اللغــة لتحديـد القسم الذي تتبع له كل كلمة في النص، ثم يتم استخدام هذا النص المصنف لتعليم خوارزميات تعلم الآلة كيفية إجراء التصنيف لنصوص جديدة غير تلـك التي تعلمت منها، ومن أشهر تقنيات تعلم الآلة المستعملة هنا: Hidden Markov Models [٧٧-٧٨] وConditional Random Fields [٧٩].

وغالبًـا ما تسمى المجموعات النصوصيــة المستعملة في تعليم خوارزميات تعلم الآلة على تصنيف أقسام الكلام والإعراب وما شابههما بـ Treebanks، ومن أشهر هذه المجموعات في الإنجليزية الـ Penn Treebank[1] ومن أشهرها في اللغة العربية Penn Arabic Treebank[2].

ومــن أدوات تصنيـف أقســام الـكلام المتاحــة للمطوريــن أداتـا AMIRA و MADA وقد سبقت الإشارة إليهما في فصول سابقة. و منها أيضًا Stanford Log-linear Part-Of-Speech Tagger[3] و الذي تصل دقته إلى ٩٦,٥٪ على النصوص العربيــة [٨٠] و يعتبر من أفضل الأدوات المتاحــة و أسهلها في الاستخدام، كما أنه يدعم لغات أخرى.

التشكيل الآلي Automatic Diacritization

وهي من الوظائف المهمة في اللغة العربية، وذلك لأن نطق أحرف اللغة العربية يختلـف باختــلاف تشكيل الحــرف، وتختلف تبعًا لذلـك المعاني والبنيــة الإعرابية للجملــة والعديـد من الخصائصــ اللغوية للنصــ. ومع ذلك فإن أغلـب النصوص العربيــة المنتشــرة في الكتب وعلى الإنترنت تخلو من التشكيل إما كليًا أو جزئيًا، أما النصوص المشكلة تشكيلًا كليًا فهي نادرة جدًا.

---

(1) http://www.cis.upenn.edu/~treebank/
(2) https://catalog.ldc.upenn.edu/LDC2003T06
(3) http://nlp.stanford.edu/software/tagger.shtml

وهـذه الظاهـرة في اللغـة العربيـة تتسبب في حالـة التباس كبـيرة تجعل من الضروري الاعتماد بشكل أكبر على السياق لفك الالتباسات. ومن أشهر التطبيقات التي يعتبر توفـر التشكيل فيها أساسيًا وضروريًا: توليـد الأصوات العربية Arabic Speech Synthesis، حيـث يكون مـن الضروري تحديد تشكيل كل حرف في الكلمة حتـى تتمكن آلة توليـد الصوت من إخـراج الصوت المناسب للحـرف. وكحل لهذه المشكلـة اقترحت بعض البحـوث بناء أنظمـة لإضافة التشكيـل الآلي، وأغلب هذه الطـرق تعتمـد على تقنيات تعلم الآلـة أو تقنيات هجينة تجمع بـين تعلم الآلة وبين القواعـد المصاغة يدويًـا، وتستخدم أغلب هذه الطـرق خصائص لفظية Lexical، وخصائـص صرفيـة Morphological، وخصائـص نحويـة Syntactic لتدريب خوارزميـات تعلم الآلة على إجراء عمليات التشكيل الآلي. ومن أمثلة الأبحاث التي تناولت الموضوع ونحيل القارئ المهتم إلى مطالعتها: [٨١-٨٣].

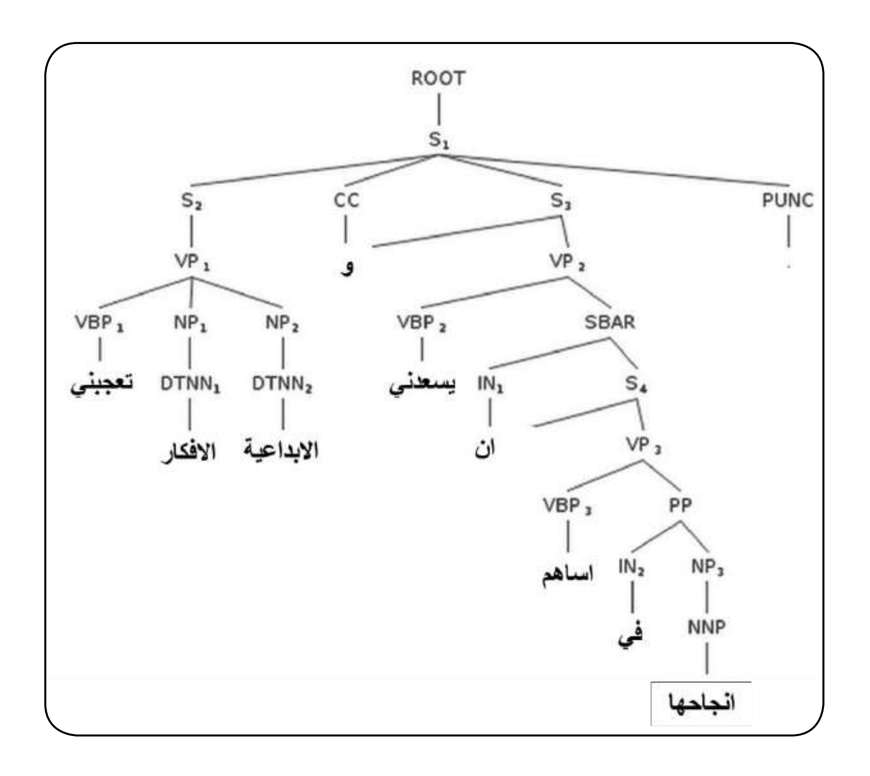
وتصل دقة أنظمـة التشكيل الآلي إلى٢,٨٨٪ وذلك في تحديد تشكيل نهايات الكلمـات كمـا في الطريقة المنشـورة في[٨٤] التي تستخدم تقنيـات التعلم العميق Deep Learning لتنفيذ المهمة.

كمـا تتضمن باقة MADA التي أشرنا إليهـا سابقًا أداة للتشكيل الآلي تصل دقتهـا في كشف حركات نهايات الكلمـات إلى ١,٨٥٪ و يمكن للباحثين و المطورين استعمالها.

تحليل البناء النحوي Syntactic Parsing

وتهـدف هذه العمليـة إلى كشف بنية الجملة، وتحديـد العلاقات النحوية بين الكلمـات والعبارات في الجملة، كتحديد الأقسـام البسيطة التي تتكون منها الجمل المركبـة وأدوات الربط بين تلك الأقسـام، مثال ذلك الجمل التي تحتوي على شرط وجـواب شـرط وأداة شرط تربط بينهمـا. وكذلك تحديد الكلمات التـي تّكوّن معًا عبارة اسمية Noun Phrase أو عبارة فعلية Verb Phrase. وغالبًا ما يتم تمثيل

هــذه العلاقــات في صــورة شجــرة جذرها في الأعلــى ويمثل كل تفرع مــن تفرعات الشجــرة قسم من الجملة، كما تمثل كل ورقــة من أوراق الشجرة النهائية كلمة من كلمــات الجملة أو علامات الترقيــم فيها (والقسم Part of speech الذي تتبع له). النقــاط الداخليــة في الشجــرة تمثل بنى نحويــة في الجملة فمثــلاً النقطةS تمثل جملــة، والنقطــة VP تمثل جملة فعلية، والنقطة SBAR تمثــل جملة تابعة كجملة جــواب الشرط أو الجملة المعطوفة. الشكل ٤ يوضح التمثيل الشجري لنتيجة تحليل البناء النحوي لجملة عربية.



شكل ٤ : التمثيل الشجري للبناء النحوي لجملة عربية

ويلاحظ أن عمليــة الإعراب النحوي تتضمن تصنيف أقسام الكلام لكل كلمة وتزيـد عليها بالذهــاب إلى مستوى أعلى من التحليل يشمـل تصنيف أقسام أشباه الجمل والتعبيرات الاسمية أو الفعلية أو أدوات الربط، وغيرها مع تمثيل هذا كله في شكل شجري يعبر عن البنية النحوية للجملة. وتسمى هذه العملية أحيانًا بالإعراب العميــق Deep Parsing وذلك بغرض التفرقة بينها وبين الإعراب السطحي Shal-low parsing الــذي يكتفــي بتحديـد أقسام الكلمــات وتحديد العبــارات الاسمية والعبارات الفعلية دون تحديد البنية الكاملة للجملة.

وتعتمـد تقنيــات الإعــراب النحوي علــى تعريف قواعـد النحــو الحاكمة للغة وتمثيلهــا بشـكل يفهمه الحاسـوب ثم تطبيــق هــذه القواعد على الجمـل لتحديد الإعــراب الصحيــح للجملــة. وتنقســم خوارزميــات الإعــراب النحــوي من حيث كيفية استخـدام قواعد النحو إلى خوارزميات تستعمـل طريقة الصعود Bottom-up وأخــرى تستعمــل طريقة النزول Top-down. الأولى تبـدأ من الكلمات (أوراق الشجــرة) وتبــدأ بتطبيق القواعــد النحوية التي تحتوي تلك الكلمــات صعودًا حتى الوصــول إلى جذر الشجــرة (الذي يمثل جملة مكتملــة). أم الطريقة الثانية فتبدأ من القواعد النحوية التي تتضمن جذر شجرة (أي تركيبة جملة كاملة) نزولًا حتى الوصول إلى الأوراق (الكلمات).

وتحتــاج خوارزميات الإعــراب أن يتم تزويدها بقواعد النحــو الخاصة باللغة حتى تتمكن من تطبيقها، ويتم تمثيل قواعد النحو حاسوبيًا على شكل مجموعة من القوانين التي تُعرِّف البُنى المسموح وجودها في الجملة. الشكل يحتوي مثال إيضاحي لبعض قواعد النحو وطريقة تمثيلها وكيفية استخدامها لإعراب جملة عربية [٨٥].

وتحتــاج عملية الإعراب إلى معرفة المعنــى المقصود كذلك وليس قواعد النحو فقط، فالجملة الواحدة – إذا ما تجاهلنا معناها المقصود– فإنها قد تحتمل أكثر من تركيب نحوي.

ولذلك فإن العديد من خوارزميات الإعراب شائعة الاستخدام حاليًا تلجأ إلى الطريقة الإحصائية Statistical Approach والتي تهدف إلى إعطاء قيمة احتمالية لـكل قاعدة مـن قواعد التركيب النحـوي بحسب مقدار تكـرر استعمالها في سياق الجملة التي يراد إعرابها، كما أنها تعطي درجة احتمالية لكل شجرة من الشجرات النحويـة العديدة التي يمكن بناؤها للجملـة الواحدة، وعادة ما تستخدم خوارزمية بحث مثل Viterbi algorithm للبحث عن أفضل شجرة في فضاء الشجرات المحتملة (دون الحاجـة لاحتساب احتمالية كل شجرة بالكامل). ويتم تعلم القيم الاحتمالية للقواعـد من خلال استخدام مجموعات نصية مصنفـة ومعربة Treebanks كتلك المستخدمة في عملية تصنيف أقسام الكلام التي أشرنا إليها في الفصل السابق.

الشجرة الإعرابية لـ "أشرب شايا لذيذا"	القواعد وشرحها باللغة العربية

root
S
VP
V NP
N ADJ
اشرب شاي ساخن

الجملة يمكن أن تكون جملة فعلية	S => VP
الجملة الفعلية تتكون من فعل وعبارة اسمية	VP => V NP
العبارة الاسمية يمكن أن تتكون من اسم وصفة	NP => N ADJ
«اشرب" فعل	V => أشرب
«شايا" اسم	N => شاي
«لذيذا" صفة	ADJ => لذيذ

شكل٥ : مثال يوضح طريقة تمثيل قواعد اللغة وتطبيقها لإعراب جملة

ومن أدوات الإعراب المتاحـة للباحثين والمطورين The Stanford Parser□□□، وهــي تدعم اللغــة العربيــة إلى جانــب الإنجليزيــة والصينيــة والإسبانية، ويمكن استخــدام نفســ الأداة لتصنيف أقسام الكلام كذلــك Part-of-speech tagging، ومن الأدوات أيضًا Bikel's Parser□□□ [٨٦] والتي تدعم اللغة العربية والإنجليزية.

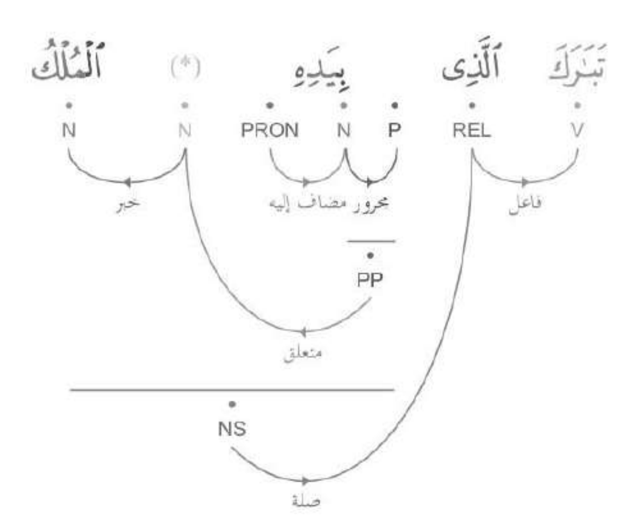
تحليل العلاقات الاعتمادية (الإعراب) Dependency Parsing

وهــو نوع آخر من التحليــل البنائي للجمل ولكنه يهــدف إلى تحديد العلاقات بــين الكلمــات وتبيان أي الكلمــات أو التعابير تعتمد معنويًا أو نحويًــا على كلمات أو تعابــير أخرى مع توضيح نوع الارتباط الذي يجمــع بينهما، ومن أمثلة ذلك ارتباط الاســم بالفعل في جملة عندما يكون الاسم فاعلًا كمــا في "شرب الطفل الحليب"، الفعــل "شــرب" مرتبط بعلاقة اعتماديــة مع الاسم "الطفل" حيــث أن الطفل هو الفاعــل، وبالمثل كلمة "الحليــب" مرتبطة بالفعل "شرب" لكونهــا المفعول به. وقد تكــون العلاقات التي تربط مكونات الجملة علاقــات معنوية Semantic أو علاقات نحوية Syntacticأو علاقة تصريفية Morphological.

ويوضــح الشكل ٦ مثال لجملــة عربية (آية قرآنية من ســورة الملك) وإعرابها والعلاقات الاعتمادية بين مكوناتها □□□.

(1) http://nlp.stanford.edu/software/lex-parser.shtml
(2) http://web.mit.edu/6.863/tools/dbparser/

□□□ المثال مأخوذ من موقع المجموعة النصية القرآنية http://corpus.quran.com



شكل ٦ : العلاقات الإعرابية والاعتمادية بين مكونات آية : تبارك الذي بيده الملك.

وقــد تناولــت العديد من الأبحاث دراسة تحليل العلاقــات الاعتمادية في اللغة العربيــة، وقد ركزت في معظمها تطبيق الخوارزميات المستخدمة في اللغات الأخرى على اللغة العربية مع إضافة خصائص صرفية وبنيوية إضافية تخص اللغة العربية حتى تساعد الخوارزميات على تحسين دقة الإعراب للجمل العربية [٨٧-٩٠].

ومن أدوات الإعراب المتاحة للباحثين والمطورين TurboParser[1] **[٩١].**

تمييز أسماء الكيانات (الأعلام) Named-Entity Recognition

كثــير من تطبيقات معالجة اللغات تتطلب أن يتمكن التطبيق من تمييز أسماء الكيانــات ومعالمتهــا بشكل خاص عنــد معالجة النص، وتشمل أسمــاء الكيانات في الغالــب أسماء الأشخاص، وأسماء الهيئات، وأسمــاء الأماكن كالدول والمحافظات والمــدن، وأحيانا تضاف إليها الأزمنة وأشيــاء أخرى. هذه هي التصنيفات الشائعة

(1) http://www.ark.cs.cmu.edu/TurboParser/

في أغلـب التطبيقات، ولكن قد تحتاج بعض المجالات إلى تحديد أصناف أخرى من أسمــاء الأعلام فمثلًا في التطبيقات الطبية تجد أصناف من مثيل أسماء الجينات وأسماء البروتينات وأسماء الأمراض وأسماء الأدوية.

شكل ٧ يعرض مثالًا لجملة عربية مأخوذة من خبر صحفي، والنتيجة المتوقعة بعد تطبيق أحد الخوارزميات العامة لتمييز أسماء الكيانات على تلك الجملة.

النص	عقدت جامعة الدول العربية اجتماعًا برئاسة أحمد العربي لمناقشة أوضاع القدس
بعد تحديد أسماء الأعلام	عقدت [جامعة الدول العربية]منظمة اجتماعًا برئاسة[أحمد العربي]شخص لمناقشة أوضاع [القدس]مكان

شكل ٧ : مثال يوضح نتيجة تطبيق خوارزمية تحديد أسماء الأعلام على نص عربي.

وتحتــاج العديــد مــن التطبيقات لهــذه العملية بشــكل أساسي حتــى لا تخلط بــين أسمــاء الكيانات التي ينبغي معاملتهــا كتعبير واحد كل الوقــت، وبين الكلمات المتفرقــة، ومن هذه التطبيقات أنظمــة اشتقاق المعلومــات Information Extraction، وأنظمــة الإجابة الآلية للأسئلة Question Answering Systems ومحركات البحث Search Engines والترجمة الآلية Machine Translation.

وكغيرها من وظائف معالجة اللغات، فــإن التقنيات المستخدمة لتمييز أسماء الكيانــات تنقسم بشــكل أساسي إلى خوارزميات تستخــدم قواعد وقوانين مصاغة يدويًــا تعــرف الحاسوب علــى الأنماط النصية التــي تظهر فيها أسمــاء الكيانات، وخوارزميــات أخرى إحصائيــة Statistical –وهي المستخدمــة الآن– وتعتمد على تقنيــات تعلم الآلــة. ومن أشهــر التقنيات المستخدمــة حاليًــا Conditional Random Fields إلا أنهــا تحتاج إلى توفــر كميات كبيرة من النصوص التي يتم تحديد أسمــاء الكيانات فيهــا يدويًا لاستخدامهــا لتدريب الخوارزمية. ومــن الطرق التي

أظهرت جدوى وهي أيضًا تحتاج إلى كمية كبيرة من النصوص -ولكن كمية الجهد اليدوي اللازم فيها أقل- تلك الطرق التراكمية التي تبدأ بمعرفة كمية قليلة من أسماء الكيانات، وتتعلم الأنماط التي تظهر فيها بشكل متكرر، ثم تستخدم هذه الأنماط للتعرف على المزيد من أسماء الكيانات، ثم منها تتعرف على مزيد من الأنماط وهكذا. وقد لخص البحث الاستقصائي المشار إليه في [٩٢] أبرز الطرق المستخدمة لتحديد أسماء الكيانات وقارن بين تلك الطرق، وعلى الرغم من كون البحث منشورًا في ٢٠٠٧ إلا أن تقنيات تحديد أسماء الكيانات كانت قد وصلت في ذلك الوقت إلى درجة عالية جدًا من الدقة، ولم تشهد أي تطور كبير خلال الفترة التالية.

أما في اللغة العربية فقد ظهر الاهتمام بهذا الموضوع متأخرًا بعض الشيء، وقد رصد الاستقصاء المشار إليه في [٩٣] أبرز ما يخص اللغة العربية في هذا الموضوع من تحديات، ولخص أهم الطرق المستخدمة فيه وعرض تقييمًا لكل منها.

وتعتبر عملية تمييز أسماء الكيانات في اللغة العربية أصعب نسبيًا من اللغات الأخرى، ومن أسباب ذلك مثلًا أن اللغة العربية لا يوجد في طريقة كتابتها أي علامة تميز أسماء الكيانات على خلاف اللغة الإنجلزية مثلًا والتي تكتب فيها أسماء الكيانات مبدوءةً بأحرف كبيرة Capital Letters، وهي خاصية من أهم الخصائص التي تعطيها خوارزميات التعلم الآلي وزنًا كبيرًا بين الخصائص اللغوية الأخرى. ومنها كذلك الثراء الصرفي للغة العربية والذي يسمح بدخول الزوائد على أسماء الكيانات وليس على الكلمات العادية فقط، مثل دخول الباء على "مكة" فتصبح "بمكة" أو دخول النا على "القدس" وإزالة أل التعريف منها فتصبح "قدسنا"، وهكذا. ومن التحديات أيضًا ندرة الموارد ذات الجودة العالية التي تحتوي قوائم بالأسماء كما يوجد في اللغات الأخرى، ومنها كذلك غياب التشكيل عن معظم النصوص العربية وينتج عنه غموض في كثير من الكلمات يجعل

عمليـة الاعتماد على السياق لحل الغموض عملية مهمة وحساسة، و تحديات أخرى مشروحة بالتفصيل في [٩٣].

ومن الطرق التي اقترحها الباحثون لتمييز أسماء الكيانات في اللغة العربية ما طـرح في [٩٤-٩٥] وهي تعتمد على مجموعة من القواعـد المصاغة يدويًا، وما هو مطروح في [٩٦-٩٨] وهي طرق تعتمد على تقنيات تعلم الآلة.

وظائف أخرى

بالإضافة إلى الوظائف التي استعرضناها في هذا الفصل هناك وظائف أخرى يتكـرر الاحتياج إليهـا في معالجة اللغات، ونذكرها هنـا أسماءها مع ترك إشارات لبعضـ المراجع المفيدة فيها. ومنها تحديد الإشـارات المتعددة للشيء الواحد Co-reference resolution (ارجـع للاستقصاء [٩٩])، وتحليل الخطاب Discourse Analysis (ارجـع لكتاب [١٠٠])، وتمييز وتقسيم موضوعات النص Topic Recognition & Segmentation [١٠١-١٠٢]، وحـل التباسـ معاني الكلمات Word Sense Disambiguation (ارجـع للاستقصاء [١٠٣])، وغيرها. وهذه الوظائف المشار إليها هنا لم تنل حظًا من الاهتمام البحثي فيما يتعلق باللغة العربية وهي بلا شك مجالات هامة ولازمة للغة العربية وتحتاج إلى من يبادر ويتصدر للبحث فيها.

٣. أشهر تطبيقات معالجة اللغات

وأخـيرا، نستعرضـ هنا أشهر تطبيقـات معالجة اللغات التـي شهدت نشاطًا بحثيًـا في اللغـة العربيـة. وتستخدم هـذه التطبيقـات في أغلب خطـوات تنفيذها الوظائف التي استعرضناها في الباب السابق.

الترجمة الآلية Machine Translation

كما ذكرنا في بداية هذا البحث، فقد كانت الترجمة الآلية من أوائل تطبيقات معالجـة اللغات التي لاقت اهتمامًا كبيرًا في الستينات وأوائل السعبينات من القرن

الماضــي، ثم خفــت الاهتمام بها بسبــب النتائج المخيبة للآمــال في ذلك الوقت، ثم عادت الأبحاث في هذا الموضوع لتنتعش في التسعينات من القرن الماضي وحتى يومنا هــذا من قرننا الحالي. ويعتبر مجــال الترجمة الآلية من المجالات البحثية الساخنة جــدًا في الوقت الحالي، وقــد حققت نجاحات وقفزات كبيرة وملموسة، ولم تعد تلك التقنيات مقصورة على البحث العلمي بل أصبحت من المنتجات المهمة في الصناعة، وتوجــد في الأســواق خدمات ترجمة آليــة تقوم على تطويرهــا والترويج لها شركات كبيرة مثل Bing Translator من مايكروسوفت وGoogle Translate وغيرها.

ويعــود الفضــل في تحقيق القفزة الكبــيرة في مجال الترجمة الآليــة إلى بروز وانتشار الطــرق الإحصائية Statistical Methods ، فبعد أن كانت عملية الترجمة تتطلــب فهمًا عميقًا لقواعد ومعاني وتراكيب اللغة المصــدر واللغة الهدف وصياغة قواعد وقوانين معقدة لترجمة النص، فإن الطرق الإحصائية أتاحت ترك استنباط هذه القوانــين والقواعد لخوارزميات التعلم الآلي، التــي تستنبط هذه القوانين من خــلال معالجة عدد كبير من المجموعات النصيــة المتوازية Parallel Corpora بعد إجــراء عملية ربط لكل جملة في المستنــد المكتوب باللغة المصدر إلى الجملة المقابلة لها في اللغة الهدف، وفعل ذلك الربط أيضًا على مستوى الكلمات. وقد سهل انتشار الانترنــت وتقنيات النشر الالكتروني كثيرًا من إمكانيــة توفير مثل هذه المجموعات النصيــة، ومــن أشهــر المصــادر لهذه المجموعــات وثائــق الأمم المتحــدة التي تتم ترجمتهــا أولًا بأول إلى كل لغات العــالم الأساسية، وكذلك ترجمات الكتب، وكذلك المواقع الالكترونية التي تحتوي صفحات متعددة اللغات، وغيرها.

وكحــال غيرها من تطبيقات معالجة اللغات فــإن الفترة الأخيرة شهدت ميلًا للجمع بين الطرق المعتمدة على القواعد المصاغة يدويًا، والطرق المعتمدة على تعلم الآلة، بحيث تصاغ بعض القواعد الضابطة التي تقيد وتحكم استخدام نتائج التعلم الآلي بهدف منع الوقوع في أخطاء يسهل ضبطها بالقوانين اللغوية.

وقـد شهدت الترجمـة الآلية من وإلى اللغـة العربية اهتمامًـا متزايدًا مؤخرًا ومن أمثلة ذلـك: ترجمة عربي-إنجليـزي [١٠٤-١٠٦] وعربي-فرنسي [١٠٧]، وعربي-صيني [١٠٨].

التلخيص الآلي Automatic Summarization

تهـدف عمليات التلخيص الآلي إلى اختصـار الكلام، وتقليل كمية النص إلى مقـدار محـدد يختاره المستخـدم أو يفرضه التطبيق مع الحفـاظ على أهم المعاني والمفاهيم الواردة في النصـ. ويفرق البحاثون بين نوعين من التلخيص الآلي: أولًا: الاختصـار الاقتباسـي Extractive Summarization وفيـه تكـون جمل الملخص مقتطعـة كما هي من نصـ المصدر، أي إن عملية التلخيصـ في هذه الحالة ما هي إلا عملية ترتيب للجمل حسب أهميتها ومقدار ما تحتويه من معلومات هامة وتغطي جوانـب الموضوع. ثانيًا: الاختصار الخلاصي Abstractive Summarization وفيه قـد تكـون جمل الملخص مولدة آليًـا أو ناتجة عن إحداث تغيـيرات وعمليات حذف أو إضافـة أو تعديل علـى الجمل القادمة من النص المصـدر، وهي أصعب بمراحل مـن التلخيصـ الاقتطاعي الاقتطاعي لأنهـا تتضمن عملية توليـد الجمل، وتحتاج إلى توفـير سبل ضمان رصانتها وتماسكها وسلامتها النحوية وصحة معانيها وعدم تناقضها مع المعلومات الواردة في المصدر.

والتلخيص الآلي -لاسيمـا التلخيص الاقتطاعي- واحد من الموضوعات التي شهدت قدرًا من الاهتمام في اللغة العربية، وجاءت أغلب الأبحاث في الموضوع لتطبق الطـرق المستخدمة في اللغات الأخرى مع إضافة بعض عمليات المعالجة التي تخص اللغة العربية، ومن أمثلة الأبحاث التي تناولت التلخيص العربي [١٠٩-١١٢].

تصنيف النصوص Text Classification

تقـوم أنظمة تصنيف النصوصـ بتحليل النص و تصنيفه بحسب موضوعه أو محتواه إلى أصناف محددة تختلف من تطبيق لآخر، ومن أمثلة ذلك تصنيف الأخبار

بحسب موضوعها إلى أخبار سياسية، أو رياضية، أو علمية، إلخ. وتعتمد أغلب هذه الأنظمة اليوم على تقنيات تعلم الآلة المخصصة لعمليات التصنيف Classification، وتستخدم في ذلك مجموعة من الخصائص من أشهرها مجموعات الكلمات n-grams المفردة والمزدوجة والثلاثية (و غالبًا يتم ذلك بعد تجذيع الكلمات Stemming)، وخصائص أخرى تختلف بحسب التطبيق.

و تصنيف النصوص العربية هو أيضًا أحد الموضوعات التي لاقت اهتمامًا من قبل عدد من الباحثين [١١٣-١١٥].

تحليل المشاعر واستكشاف الآراء Sentiment Analysis

وهي عملية يتم فيها تحليل المحتوى النصي بهدف الكشف عما يحمله من مشاعر سلبية أو إيجابية أو محايدة، وتستخدم هذه العمليات بشكل كبير في تحليل المراجعات التي يكتبها الناس على الإنترنت تعليقًا على ما يشترونه من منتجات، أو تلك التي يعلق بها الناس على ما يشاهدونه من أفلام مصورة أو مقاطع صوتية، وكذلك في قياس الرأي العام تجاه القضايا العامة والقضايا السياسية، وتفيد هذه العمليات صناع القرار في الشركات أو الدوائر الرسمية لاتخاذ ما يلزم من قرارات وسياسات. وقد ازداد الاهتمام بهذا الموضوع بشكل كبير مع انتشار الانترنت، ومع ثورة شبكات التواصل الاجتماعي تحديداً، حيث أصبحت هذه الشبكات مليئة بالنص المحمل بالمشاعر والآراء التي يمكن لصناع القرار الوصول إليها وتحليلها والخروج باستنتاجات من خلالها.

وتعتمد هذه الأنظمة على تحديد الكلمات والتعبيرات السلبية والإيجابية في اللغة فمثلاً كلمة "سعيد" هي كلمة إيجابية وفي المقابل كلمة "حزين" هي كلمة سلبية. وقد تعتمد قطبية الكلمات على الموضوع فمثلًا عندما يوصف سيناريو فيلم ما بأنه "غير متوقع" فهذا قد يكون مؤشر إيجابي يدل على أنه مشوق، في المقابل عندما يوصف مقود التحكم في السيارة بأن "غير متوقع" فهذا تعبير سلبي يدل على أنه خطير ويصعب التحكم فيه بدقة عند القيادة.

وقـد ظهر اهتمام كبير مؤخرًا بتطبيقات تحليل المشاعر واستكشاف الآراء في اللغة العربية وظهرت العديد من البحوث التي اقترحت آليات لبناء قوائم بالتعابير الإيجابية والسلبية المستخدمة في اللغة العربية [١١٦-١١٧]، وأخرى تناولت تحليل نصوص شبكات التواصل الاجتماعي كمنتديات الحوار ومواقع التدوين القصير لقياس الرأي العام [١١٨-١٢٠].

خاتمة

عرّفنـا في هـذا البحـث الاستقصائـي بتقنيـات معالجـة اللغـات الطبيعية، واستعرضنـا أهـم الوظائف الأساسية التي يلزم إجراؤهـا على النصوص في أغلب تطبيقات معالجة اللغات، وأوضحنا ما للغة العربية من خصوصيات في كل من تلك الوظائـف مع ترك إشارات مرجعية للبحوث التي تناولت كل وظيفة حتى يعود إليها القارئ المهتم بمعرفـة التفاصيل. كما أشرنا إلى العديد من الوظائف الأخرى التي لم يتطـرق إليهـا البحـث في اللغة العربية فيمـا نعلم وتحتاج إلى مبـادرة الباحثين لطـرق بابها. ثـم عرضنا نماذج لتطبيقـات مهمة لتقنيات معالجـة اللغات لاسيما تلك التي اهتم بهـا الباحثون في اللغة العربية، وتركنا كذلك إشارات مرجعية لتلك الأبحاث التي يمكن أن يرجع إليها القارئ المهتم بالتفاصيل.

المراجع

1. S. J. a. N. P. Russell، Artificial Intelligence: A Modern Approach، Pearson Education، 2003.

2. A. M. Turing، "Computing machinery and intelligence،" Mind، pp. 433-460، 1950.

3. A. C. I. a. A. V. Saygin، "Turing Test: 50 Years Later،" Minds and Machines ، vol. 10، no. 4، 2000.

.4 701"Translator،" IBM، 8 January 1954. [Online]. Available: http://www-03.ibm.com/ibm/history/exhibits/701/701__translator.html. [Accessed 3 December 2014.[

5. J. Hutchins، "The first public demonstration of machine translation: the Georgetown-IBM system، 7th January 1954،" in AMTA conference، 2004.

6. W. Lehnert، "WHAT MAKES SAM RUN? SCRIPT BASED TECHNIQUES FOR QUESTION ANSWERING،" in Theoretical Issues in Natural Language Processing: Supplement، 1975.

7. K. R. McKeown، "Paraphrasing Using Given and New Information in a Question-Answer System،" in 17th Annual Meeting of the Association for Computational Linguistics، 1979.

8. L. Karttunen، "DISCOURSE REFERENTS،" in INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS COLING 1969: Preprint No. 60، 1969.

9. J. H. Clippinger، "SPEAKING WITH MANYTONGUES: SOME PROBLEMS IN MODELING SPEAKERS OF ACTUAL DISCOURSE،" in Theoretical Issues in Natural Language Processing، 1975.

10. S. K. a. R. Simmons، "A computational approach to grammatical coding of English Words،" JACM، vol. 10، p. 334–337، 1963.

11. B. B. G. a. G. M. Rubin، "Automated grammatical tagging of English،" Department of Linguistics، Brown University، 1971.

12. K. W. Church، "A stochastic parts program and noun phrase parser for unrestricted text،" in The second conference on Applied natural language processing، 1988.

13. S. J. DeRose، "Grammatical category disambiguation by statistical optimization،" Computational Linguistics، vol. 14، no. 1، pp. 31-39، 1988.

14. J. Tait، "Automatic summarization of english texts،" PhD thesis، University of Cambridge، Cambridge، UK، 1983.

15. G. a. M. M. J. Salton، Introduction to modern information retrieval، McGill ، 1983.

16. P. Turney، "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews،" in The 40th Annual Meeting of the Association for Computational Linguistics، Philadelphia، Pennsylvania، 2002.

17. L. L. a. S. V. Bo Pang، "Thumbs up? Sentiment classification using machine learning techniques،" in EMNLP، 2002.

18. R. Al-Shalabi and M. Evens، "A Computational Morphology System for Arabic،" in Workshop on Computational Approaches to Semitic Languages، 1998.

19. M. Smets، "Paradigmatic Treatment of Arabic Morphology،" in Workshop on Computational Approaches to Semitic Languages، 1998.

20. K. R. Beesley، "Arabic Finite-State Morphological Analysis and Generation،" in COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics، 1996.

21. K. R. Beesley، "Consonant Spreading in Arabic Stems،" in 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics، Volume 1، 1998.

22. H. El-Shishiny، "A Formal Description of Arabic Syntax in Definite Clause Grammar،" in COLNG 1990 Volume

3، Papers presented to the 13th International Conference on Computational Linguistics، 1990.

23. R. a. M. E. a. D. J. a. S. D. a. M. S. a. S. R. a. M. J. a. Z. F. O. a. C.-B. C. Zbib، "Machine Translation of Arabic Dialects،" in Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics، Human Language Technologies، 2012.

24. W. Salloum and N. Habash، "Elissa، A Dialectal to Standard Arabic Machine Translation System،" in Proceedings of COLING 2012، Demonstration Papers.

25. A. Ittycheriah and S. Roukos، "A Maximum Entropy Word Aligner for Arabic-English Machine Translation،" in Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing، 2005.

26. A. Azmi، "Ikhtasir — A user selected compression ratio Arabic text summarization system،" in International Conference on Natural Language Processing and Knowledge Engineering، 2009. NLP-KE 2009.، 2009.

27. J. D. Schlesinger، D. P. O'Leary and J. M. Conroy، "Arabic/English Multi-document Summarization with CLASSY - The Past and the Future،" Computational Linguistics and Intelligent Text Processing، pp. 568-581، 2008.

28. G. Badaro، R. Baly، H. Hajj، N. Habash and W. El-Hajj، "A Large Scale Arabic Sentiment Lexicon for Arabic Opinion Mining،" in Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)، 2014.

29. A. Abu-Jbara، B. King، M. Diab and D. Radev، "Identifying Opinion Subgroups in Arabic Online Discussions،" in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)، 2013.

30. D. Vergyri and K. Kirchhoff، "Automatic Diacritization of Arabic for Acoustic Modeling in Speech Recognition،" in Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages، 2004.

31. D. Vergyri and K. Kirchhoff، "Arabic Diacritization Using Weighted Finite-State Transducers،" in Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages، 2004.

32. K. Shaalan، H. M. Abo Bakr and I. Ziedan، "A Hybrid Approach for Building Arabic Diacritizer،" in Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages، 2009.

33. N. Habash، O. Rambow and G. Kiraz، "Morphological Analysis and Generation for Arabic Dialects،" in Pro-

ceedings of the ACL Workshop on Computational Approaches to Semitic Languages، 2005.

34. F. O. Zaidan and C. Callison-Burch، "Arabic Dialect Identification،" Computational Linguistics، vol. 40، no. 1، 2014.

35. E. Mohamed، B. Mohit and K. Oflazer، "Transforming Standard Arabic to Colloquial Arabic،" in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)، 2012.

36. N. Habash، M. Diab and O. Rambow، "Conventional Orthography for Dialectal Arabic،" in Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)، 2012.

37. N. Habash، Introduction To Arabic Natural Language Processing، 2010.

38. A. Farghaly and K. Shaalan، "Arabic Natural Language Processing: Challenges and Solutions،" ACM Transactions on Asian Language Information Processing (TALIP)، vol. 8، no. 4، 2009.

39. A. a. H. N. El Kholy، "Orthographic and morphological processing for English--Arabic statistical machine translation،" Machine Translation، vol. 26، pp. 25-45، 2012.

40. Y. a. I. F. Elarian، "A Lexicon of Connected Components for Arabic Optical Text Recognition،" in First International Workshop on Frontiers in Arabic Handwriting Recognition، Istanbul، Turkey، 2011.

41. T. Buckwalter، "Buckwalter Arabic Morphological Analyzer Version 1.0،" Linguistic Data Consortium (LDC)، 2002.

42. N. a. S. A. a. B. T. Habash، "On Arabic Transliteration،" in Arabic computational morphology، Springer، 2007، pp. 15-22.

43. A. Roberts، "Buckwalter2Unicode،" 15 September 2004. [Online]. Available: http://www.redhat.com/archives/fedora-extras-commits/2007-June/msg03617.html. [Accessed 4 12 2014.[

44. T. Q. A. Corpus، "Java API - Buckwalter Transliteration،" [Online]. Available: http://corpus.quran.com/java/buckwalter.jsp. [Accessed 5 12 2014.[

45. M. A. Attia، "Arabic Tokenization System،" in The 5th Workshop on Important Unresolved Matters، Prague، Czech Republic، 2007.

46. N. a. S. F. Habash، "Arabic preprocessing schemes for statistical machine translation،" in NAACL-HLT، New York City، NY، USA، 2006.

47. N. a. R. O. a. R. R. Habash، "Mada+ tokan: A toolkit for arabic tokenization، diacritization، morphological disambiguation، pos tagging، stemming and lemmatization،" in The 2nd International Conference on Arabic Language Resources and Tools (MEDAR)، Cairo، Egypt، 2009.

48. N. a. R. O. a. R. R. Habash، "Morphological Tagging for Arabic،" [Online]. Available: http://www1.cs.columbia.edu/~rambow/software-downloads/MADA__Distribution.html. [Accessed 5 12 2014.[

49. M. Diab، "Second generation AMIRA tools for Arabic processing: Fast and robust tokenization، POS tagging، and base phrase chunking،" in 2nd International Conference on Arabic Language Resources and Tools، 2009.

50. M. Diab، "AMIRA 2.1،" 25 10 2011. [Online]. Available: http://www.ibridgenetwork.org/columbia/ir__ms-12s-2-1. [Accessed 5 12 2014.[

51. S. R. Anderson، A-morphous morphology، Cambridge University Press، 1992.

52. I. A. a. A.-K. I. A. Al-Sughaiyer، "Arabic morphological analysis techniques: A comprehensive survey،" Journal of the American Society for Information Science and Technology، vol. 55، no. 3، pp. 189-213، 2004.

53. K. a. H. N. Dukes، "Morphological Annotation of Quranic Arabic،" in European Languages Resources Association (ELRA)، Valletta، Malta، 2010.

54. N. a. E. R. a. H. A. Habash، "A Morphological Analyzer for Egyptian Arabic،" in Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology، Montreal، Canada، 2012.

55. M. El-Affendi، "An algebraic algorithm for Arabic morphological analysis،" The Arabian Journal for Science and Engineering، vol. 16، no. 4B، p. 605– 611، 1991.

56. S. a. A.-A. F. Al-Fedaghi، "A new algorithm to generate root-pattern forms،" in In Proceedings of the 11th National Computer Conference ، Dhahran، Saudi Arabia، 1989.

57. R. Al-Shalabi، "Design and implementation of an Arabic morphological system to support natural language processing،" Doctoral dissertation، Illinois Institute of Technology، Chicago، IL، 1996.

58. I. a. A.-S. I. Al-Kharashi، "Rule merging in a rule based Arabic stemmer،" in In Proceedings of the 19th International Conference on Computational Linguistics (COLING-02)، Taipei، Taiwan، 2002.

59. T. Buckwalter، "Buckwalter Arabic Morphological Analyzer Version 2.0،" Linguistic Data Consortium، 2004. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2004L02. [Accessed 5 12 2014.[

60. D. G. B. B. S. K. A. B. S. K. Mohamed Maamouri، "LDC Standard Arabic Morphological Analyzer (SAMA) Version 3.1،" Linguistic Data Consortium، 2010. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2010L01. [Accessed 5 12 2014.[

61. O. a. B. V. a. B. T. Smrž، "ElixirFM،" [Online]. Available: http://quest.ms.mff.cuni.cz/cgi-bin/elixir/index.fcgi?mode=home. [Accessed 5 12 2014.[

62. O. a. B. V. a. B. T. Smrž، "ElixirFM: Functional Arabic Morphology،" 3 7 2013. [Online]. Available: http://sourceforge.net/projects/elixir-fm/. [Accessed 6 12 2014.[

63. D. A. Hull، "Stemming algorithms: A case study for detailed evaluation،" JASIS، vol. 47، no. 1، pp. 70-84، 1996.

64. W. a. P. R. Kraaij، "Viewing stemming as recall enhancement.،" in ACM SIGIR96، 1996.

65. P. Willett، "The Porter stemming algorithm: then and now،" Electronic library and information systems، vol. 40، no. 3، pp. 219-223، 2006.

66. F. C. L. M. F. a. W. P. Ekmekcioglu، "Stemming and n-gram matching for term conflation in Turkish texts،" Information Research News، vol. 7، no. 1، pp. 2-6، 1996.

.76 "Corpus-based stemming using cooccurrence of word variants،" ACM Transactions on Information Systems، vol. 16، no. 1، pp. 61-81، 1998.

68. I. A. a. E. M. W. Al-Kharashi، "Comparing words، stems، and roots as index terms in an Arabic information retrieval system،" Journal of the American Society for Information Science، vol. 45، no. 8، pp. 548-560، 1994.

69. T. Buckwalter، "Qamus: Arabic lexicography،" [Online]. Available: http://www.qamus.org./

70. A. N. a. A.-F. De Roeck، "A morphologically senstive clustering algorithm for identifying Arabic roots،" in ACL 2000، Hong Kong، 2000.

71. L. S. a. B. L. a. C. M. E. Larkey، "Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis،" in 25th annual international ACM SIGIR conference on Research and development in information retrieval، 2002.

72. S. Khoja، "Arabic stemmer،" [Online]. Available: http://zeus.cs.pacificu.edu/shereen/research.htm#stemming. [Accessed 6 12 2014.[

73. M. a. H. K. a. J. D. Diab، "Automatic tagging of Arabic text: From raw text to base phrase chunks،" in HLT-NAACL 2004: Short Papers، 2004.

74. N. a. R. R. M. Habash، "Catib: The columbia arabic treebank،" in The ACL-IJCNLP 2009 Conference Short Papers، 2009.

75. S. a. G. R. a. K. G. Khoja، "An Arabic Tagset for the Morphosyntactic Tagging of Arabic،" Lancaster University، 2001.

76. O. a. Z. P. Smrz، "Sherds from an arabic treebanking mosaic،" Prague Bulletin of Mathematical Linguistics، 2002.

77. K. W. Church، "A stochastic parts program and noun phrase parser for unrestricted text،" in The second conference on Applied natural language processing، Stroudsburg، PA، 1988.

78. S. J. DeRose، "Grammatical category disambiguation by statistical optimization،" Computational Linguistics، vol. 14، no. 1، p. 31–39، 1988.

79. J. a. M. A. a. P. F. C. Lafferty، "Conditional random fields: Probabilistic models for segmenting and labeling sequence data،" in 18th International Conference on Machine Learning، 2001.

80. K. a. K. D. a. M. C. D. a. S. Y. Toutanova، "Feature-rich part-of-speech tagging with a cyclic dependency net-

work." in the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology، 2003.

81. I. Zitouni، J. S. Sorensen and R. Sarikaya، "Maximum Entropy Based Restoration of Arabic Diacritics." in Workshop on Computational Approaches to Semitic Languages، Sydney-Australia، 2006.

82. K. a. E. G. Adali، "Vowel and Diacritic Restoration for Social Media Texts." in Workshop on Language Analysis for Social Media (LASM) ، Gothenburg، Sweden، 2014.

83. A. a. R. M. a. M. R. H. a. R. A. Al Sallab، "Automatic Arabic diacritics restoration based on deep nets." in EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)، Doha، Qatar، 2014.

84. M. A. a. A. S. A. A. a. R. H. M. a. R. A. Rashwan، "Automatic Arabic diacritics restoration based on deep nets." in ANLP، 2014.

85. S. a. M. H. a. A. M. S. Alqrainy، "Context-Free Grammar Analysis for Arabic Sentences." International Journal of Computer Applications (0975 - 8887)، vol. 53، no. 3، pp. 7 - 11، 2012.

86. D. M. Bikel، "Intricacies of Collins' parsing model." Computational Linguistics، vol. 30، p. 479–511، 2004.

87. Y. a. H. N. a. R. O. Marton، "Dependency Parsing of Modern Standard Arabic with Lexical and Inflectional Features،" Computational Linguistics، 2013.

88. S. Tratz، "A Cross-Task Flexible Transition Model for Arabic Tokenization، Affix Detection، Affix Labeling، POS Tagging، and Dependency Parsing،" in The Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages، 2013.

89. E. Mohamed، "The Effect of Automatic Tokenization، Vocalization، Stemming، and POS Tagging on Arabic Dependency Parsing،" in The Fifteenth Conference on Computational Natural Language Learning، Portland، Oregon، USA، 2011.

90. a. A. E. a. M. S. A.-B. Dukes، "Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank،" in the Seventh conference on International Language Resources and Evaluation (LREC›10)، Valletta، Malta، 2010.

91. A. F. a. S. N. A. a. X. E. P. Martins، "Concise integer linear programming formulations for dependency parsing،" in The Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP، 2009.

92. D. a. S. S. Nadeau، "A survey of named entity recognition and classification،" Lingvisticae Investigationes، vol. 30، no. 1، pp. 3 - 26، 2007.

93. K. Shaalan، "A Survey of Arabic Named Entity Recognition and Classification،" Computational Linguistics، vol. 40، no. 2، 2013.

94. K. a. H. R. Shaalan، "Person name entity recognition for Arabic،" in Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources، Semitic 2007، Stroudsburg، PA، 2007.

95. K. a. H. R. Shaalan، "Arabic named entity recognition from diverse text types،" Advances in Natural Language Processing، vol. 5221، 2008.

96. Y. M. D. a. P. R. Benajiba، "Arabic named entity recognition: An SVM-based approach،" in Arab International Conference on Information Technology، Hammamet، 2008.

97. Y. M. D. a. P. R. Benajiba، "Arabic named entity recognition using optimized feature sets،" in Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)، Stroudsburg، PA، 2008.

98. Y. a. P. R. 2. Benajiba، "ANERsys 2.0: Conquering the NER task for the Arabic language by combining the maximum entropy with POS-tag Information،" in

Workshop on Natural Language-Independent Engineering، 3rd Indian International Conference on Artificial Intelligence، Mumbay، 2007.

99. P. Elango، "Coreference resolution: A survey،" University of Wisconsin، Madison، WI، 2005.

100. D. a. T. D. a. H. H. E. Schiffrin، The handbook of discourse analysis، John Wiley & Sons، 2008.

101. J. a. R. C. Arguello، "Topic segmentation of dialogue،" in HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech، 2006.

102. J. C. Reynar، "Topic segmentation: Algorithms and applications،" University of Pennsylvania، 1998.

103. R. Navigli، "Word sense disambiguation: A survey،" ACM Computing Surveys (CSUR)، vol. 41، no. 2، 2009.

104. I. a. Z. R. a. G. J. Badr، "Segmentation for English-to-Arabic statistical machine translation،" in 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers، 2008.

105. I. a. Z. R. a. G. J. Badr، "Syntactic phrase reordering for English-to-Arabic statistical machine translation،" in 12th Conference of the European Chapter of the Association for Computational Linguistics، 2009.

106. H. a. L. A. Al-Haj، "The impact of Arabic morphological segmentation on broad-coverage English-to-Arabic statistical machine translation،" Machine translation، vol. 26، 2012.

107. S. a. E. I. A. a. N. H. Hasan، "Creating a large-scale Arabic to French statistical machine translation system،" in Fifth Int. Conf. on Language Resources and Evaluation (LREC)، 2006.

108. N. a. H. J. Habash، "Improving Arabic-Chinese statistical machine translation using English as pivot language،" in Proceedings of the Fourth Workshop on Statistical Machine Translation، 2009.

109. H. a. N. O. a. B. P. Oufaida، "Multilingual Summarization Experiments on English، Arabic and French (R}\'e{sum}\'e{ Automatique Multilingue Exp}\'e{rimentations sur l'Anglais، l'Arabe et le Fran}\c}c{{ais) [in French،" in Proceedings of TALN 2014 (Volume 2: Short Papers)، Marseille، France، 2014.

110. L. a. F. C. a. E.-H. M. a. G. G. Li، "ulti-document multilingual summarization corpus preparation، Part 1: Arabic، English، Greek، Chinese، Romanian،" in MultiLing 2013 Workshop on Multilingual Multi-document Summarization، Sofia، Bulgaria، 2013.

111. F. S. a. L. G. Douzidia، "Lakhas، an Arabic summarization system،" in DUC2004، 2004.

112. J. D. a. O. D. P. a. C. J. M. Schlesinger، "Arabic/English multi-document summarization with CLASSY—the past and the future،" in Computational Linguistics and Intelligent Text Processing، Springer، 2008، pp. 568--581.

113. L. Khreisat، "Arabic text classification using N-gram frequency statistics a comparative study،" in Conference on Data Mining| DMIN›06، 2006.

114. A. El-Halees، "Arabic text classification using maximum entropy،" The Islamic University Journal، vol. 15، no. 1، pp. 157--167، 2007.

115. S. A. A. A.-T. ،. A. K. M. S. a. A.-R. Al-Harbi، "Automatic Arabic text classification،" in The 9th International Conference on the Statistical Analysis of Textual Data، Lyon، France، 2008.

116. G. a. B. R. a. H. H. a. H. N. a. E.-H. W. Badaro، "A Large Scale Arabic Sentiment Lexicon for Arabic Opinion Mining،" in The EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)، Doha، Qatar، 2014.

117. M. a. D. M. Abdul-Mageed، "SANA: A Large Scale Multi-Genre، Multi-Dialect Lexicon for Arabic Subjectivity and Sentiment Analysis،" in The Ninth International Conference on Language Resources and Evaluation (LREC-2014)، Reykjavik، Iceland، 2014.

118. E. a. R. V. Refaee، "Evaluating Distant Supervision for Subjectivity and Sentiment Analysis on Arabic Twitter Feeds،" in EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)، Doha، Qatar، 2014.

119. A. a. K. B. a. D. M. a. R. D. Abu-Jbara، "Identifying Opinion Subgroups in Arabic Online Discussions،" in the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)، Sofia، Bulgaria، 2013.

120. A. a. D. K. Mourad، "Subjectivity and Sentiment Analysis of Modern Standard Arabic and Arabic Microblogs،" in the 4th Workshop on Computational Approaches to Subjectivity، Sentiment and Social Media Analysis، Atlanta، Georgia، 2013.

.121 "ARNE: A tool for named entity recognition from Arabic text،" in Fourth Workshop on Computational Approaches to Arabic Script-based Languages (CAASL4)، San Diego، CA، 2012.

122. A. Andreyewsky، "SUBCLASSIFICATION OF PARTS OF SPEECH IN RUSSIAN: VERBS،" in COLING، 1965.

استخدام القواعد الاستنباطية في تحسين أداء التشكيل الآلي

أ. عمر السيد شعبان (*)

omar.s.shaaban@gmail.com

(*) باحث في مجال الحوسبة العربية، درس الهندسة البرمجية في جامعة الملك فهد للبترول والمعادن وحصل منها على درجة البكلوريوس عام ٢٠٠٩. حصل على الماجستير من نفس الجامعة عام ٢٠١٣ في تخصص علوم الحاسب الآلي، وكانت رسالة الماجستير عن التشكيل الآلي في اللغة العربية. من اهتماماته الأخرى التعرف الآلي على الكلام، والتصحيح الآلي للنصوص.

ملخص[1]

يُعتبر التشكيل الآلي في اللغة العربية أحد أهم المباحث في مجال الحوسبة العربية، حيث يُمثل مرتكزًا رئيسًا لمباحث أخرى في نفس المجال كالتعرف الآلي على الكلام العربي، والنُّطق الآلي، وفي مرحلة تحليل كلام اللغة المصدر للترجمة الآلية. كما يُمكن الاستفادة منه من قبل دارسي اللغة العربية حيث يُعتبر غموض معنى الكلمة غير المُشكَّلة تحديًّا كبيرًا لديهم وخاصة المبتدئين منهم.

قام الباحثون باتباع طرق عديدة في التشكيل الآلي أغلبها وأنجحها يعتمد على الطرق الإحصائية والتعلم الآلي. كما قام بعضهم بمحاولة اتباع الطرق النحوية التي تعتمد على قواعد اللغة العربية وإن كان نجاحهم في ذلك مقصورًا على الجمل القصيرة البسيطة. وربما كان السبب الرئيس في ذلك هو الغنى اللغوي الذي تتميز به اللغة العربية عن بقية اللغات.

في هذا البحث نُقدم طريقة تعتمد على استنباط قواعد سياقية وحرفية عن طريق استخراجها من مكانز اللغة العربية المشكَّلة يدويًّا. وقد قمنا باستخدام تلك القواعد في مرحلتين: مرحلة التنظيف وهي المرحلة التي يتم فيها استبعاد التشكيل حسب درجة نُدرته، ومرحلة التطبيق وفيها يتم تطبيق هذه القواعد على الأحرف غير المُشكَّلة في النص.

وتظهر نتائج التجارب التي قمنا بها تحسنًا كبيرًا في نسبة التشكيل تصل إلى ١١٪ في المتوسط، وكذلك تحسنًا في نسبة الخطأ الكلمي بمتوسط ٧٪، وفي نسبة الخطأ الحرفي بمتوسط ١٠٪.

[1] يعتمد هذا البحث على رسالة بعنوان "استرجاع التشكيل آليًّا للنصوص العربية" قدمت لجامعة الملك فهد للبترول والمعادن ومنح بموجبها الباحث درجة الماجستير تحت إشراف د.حسني المحتسب.

مقدمة

تعـد معالجة اللغة الطبيعيـة (Natural Language Processing NLP) أحد المواضيع المهمة في علوم الحاسـوب واللسانيات الحاسوبية، وقد تطورت كثيرا على مدار العقود الثلاثة الماضية. وتكمن أهمية معالجة اللغة العربية في انتشار الإنترنت والأجهزة المحمولة التي تتطلب المزيد من التفاعل الطبيعي مع المستخدمين. فبعض تطبيقات معالجة اللغة الطبيعية مخصصة لأهداف لغوية مستقلة كالمدقق والمصحح الإملائـي والنحوي وبعضها تستخدم كأدوات مساعدة لتحقيق أغراض أكثر تعقيدًا كالتعـرف الآلي علـى الـكلام (Automatic Speech Recognition ASR) حيـث يستطيع الشخص إطلاق أوامر صوتية أو إملاء نصٍّ للحاسوب، وكذلك مثل تحويل النصـ إلى كلام (Text-to-Speech TTS) حيث يقوم الحاسوب بنطق نص ما أو إعلام المستخدم بأمر ما.

التشكيل في الكتابة العربية

تتكون الأبجديـة العربية من ٢٨ حرفًا، منها ٢٥ حرف صامت، و٣ أحرف علة (صائتـة) (الألـف، والـواو، والياء (، ويمكن للـواو والياء أن يكونـا ساكنين (شبه صائتة) كذلك [١].

يمكـن لكل حرف صامـت أن يُشكل بإحدى ١٤ طريقة كمـا يبينها الجدول ١، ويمكـن تصنيف هذه الطرق في خمس مجموعـات: الأولى تمثل الحركات البسيطة وهي الفتحة والضمة والكسرة، والثانية تمثل حركتي التقطيع وهما السكون والشدة، والثالثـة تمثـل التنويـن (تنوين فتـح، تنوين ضم، تنويـن كسر)، أمـا المجموعتين الرابعـة والخامسـة فتمثـلان الحـركات المركبـة مـن الشـدة مـع حركـة مـن الحركات البسيطة.

يمكـن أن نقسـم عملية تشكيـل الكلمة في اللغـة العربية إلى قسمـين: تشكيل صرفي، وتشكيل إعرابي. ونعني بالتشكيل الصرفي التشكيل الداخلي للكلمة الذي لا يتغـير بتغير موقع الكلمة في الجملـة. أما التشكيل الإعرابي فيكون عادة في الحرف الأخـير مـن الكلمة ويمكـن أن يتغير بتغير موقـع الكلمة في الجملـة. ويمكن اعتبار القسـم الأول العامل الأساس في تحديد معنى الكلمة، أما القسم الثاني فهو العامل الأساس في تحديد معنى الجملة ككل. ويمكن عند التشكيل الآلي، استخدام الطرق الإحصائيـة لاستنتـاج القسـم الأول، أمـا القسم الثـاني فلا يمكـن استنتاجه إلا باستخدام القواعد النحوية.

جدول ١ تقسيم التشكيل العربي

الرقم	القسم	الحركة	مثال
١	الحركات البسيطة	الفتحة (ـَ)	بَ
٢		الضمة (ـُ)	بِ
٣		الكسرة (ـِ)	بُ
٤	حركات التقطيع	السكون (ـْ)	بْ
٥		الشدة (ـّ)	بّ
٦	التنوين	تنوين الفتح (ـً)	بً
٧		تنوين الضم (ـٌ)	بٌ
٨		تنوين الكسر (ـٍ)	بٍ
٩	الشدة + الحركات البسيطة	الشدة + الفتحة (ـَّ)	بَّ
١٠		الشدة + الضمة (ـُّ)	بُّ
١١		الشدة + الكسرة (ـِّ)	بِّ
١٢	الشدة + التنوين	الشدة + (ـًّ)	بًّ
١٣		الشدة + تنوين الضم (ـٌّ)	بٌّ
١٤		الشدة + تنوين الكسر (ـٍّ)	بٍّ

تعريف المسألة

نهدف في هذا البحث إلى تطبيق التشكيل الآلي على النصوص العربية، وبمعنى أدق؛ يمكن وصف مسألة التشكيل الآلي بأنها عملية استرجاع التشكيل الناقص من النصوص غير المشكلة (أو المشكلة جزئيًا). يبين مثالا على نص غير مشكل (أ)، والتشكيل المفترض بعد تطبيق التشكيل الآلي (ب).

خير الناس أنفعهم للناس	خَيْرُ النَّاسِ أَنْفَعُهُمْ لِلنَّاسِ
(أ) المدخل	(ب) المخرج

رسم توضيحي ١ : مثال على مخرجات عملية التشكيل الآلي

التشكيل الآلي هو أحد قضايا معالجة اللغة الطبيعية التي يمكن اعتبارها مستقلة لها تطبيقاتها الخاصة، كما يمكن اعتبارها أداة مساعدة لتطبيقات أكثر تعقيدًا. وتظهر هذه المسألة في اللغات السامية بشكل خاص كالعربية والعبرية، وكذلك في بعض اللغات الأخرى مثل اليونانية والكورية [٢]. وباعتبار اللغة العربية يُحذف التشكيل عادة تاركًا مهمة تحديد معنى الجملة لفطنة القارئ، وهذه ليست مشكلة للقارئ الخبير الذي يستطيع بخبرته تحديد المعنى المقصود بناء على السياق، ولكن تبرز مشكلة تقدير التشكيل للقارئ المبتدئ – وكذلك للحاسوب – حيث لا يمكن تحديد معنى الجملة بسهولة. ومن هذا المنطلق تظهر الحاجة إلى التشكيل الآلي.

تطبيقات التشكيل الآلي

يمكن استخدام التشكيل الآلي بصورة مستقلة أو بصورة مساعدة لمسائل أخرى، فعند تطبيقه يمكن أن يقلل من غموض النص ويساعد في تحديد المعنى المقصود، وهذا مفيد بحد ذاته. كذلك فإن تشكيل النص هو أمر أساس في مسألتي التعرف الآلي على الكلام وتحويل النص إلى كلام والترجمة الآلية. إذ تُستخدم

الطرق الإحصائية في التعرف على الكلام على الأغلب والتي تعتمد على وجود مكنز مُشكَّل، ولأجل الحصول على هذا المكنز فإن جهود الباحثين تستنزف في بنائه بصورة يدوية، وهذا أمر شاق للغاية. أما عند تحويل النص إلى كلام فيتطلب أن يكون النص المراد نطقه مشكلا تشكيلا كاملا. ولذا، فإن بناء نظام للتشكيل الآلي إضافة ضرورية لتطبيقات وبحوث الصوتيات العربية.

أساليب التشكيل الآلي

لقد ازدادت الأبحاث المنشورة المتعلقة بالتشكيل الآلي، سواء للعربية أو غيرها، في العقد الأخير، وفي هذا الفصل نعرض آخر الأبحاث في هذا الموضوع. إذ نعرض الأبحاث المتعلقة باللغة العربية في القسم ١,٢، بينما نعرض بعض الأبحاث المتعلقة باللغات الأخرى في القسم ٢,٢. أما القسم ٣,٢ فيعرض مقارنة بين الطرق المستخدمة في الأبحاث التي اطلعنا عليها.

التشكيل الآلي في اللغة العربية

قام رشوان وآخرون [١] بعرض نظام ستوكاستيكي للتشكيل الآلي ذي الوضعين المزدوجين، حيث يقوم الوضع الأول بالبحث في قاموس يحتوي على كلمات كاملة ومشكلّة، باستخدام البحث الشعري A×، وتقدير الاحتمال التسلسلي طويل المدى لأجل الحصول على أقرب تشكيل ممكن. وفي حال لم تكن الكلمة (أو سلسلة الكلمات) موجودة في القاموس (Out-Of-Vocabulary OOV)؛ يقوم الوضع الثاني بتحليل كل كلمة إلى جميع احتمالاتها الصرفية ثم يعيد البحث في القاموس لكل احتمال. يقول المؤلفون بأن النظام حقق نسبة خطأ ٣,١٪ على مستوى الكلمات، و١٢,٥٪ على مستوى الأحرف.

واتبع الزيتوني وساريكايا [٣] طريقة الإنتروبيا القصوى (Maximum En-tropy) لاستنتاج التشكيل. وفي هذه الطريقة يمكن استخدام معلوماتٍ متعددة بصورة تكاملية كالخصائص الكلمية، والقطعية، ووسوم أجزاء الكلام. وعلى هذا؛

قـام المؤلفان بتعريف مسألة التشكيـل الآلي على أنها مسألـة تصنيف، واستخدما مصَنِّـف MaxEnt، كما استخدما في تجاربهما المكنـز LDC Arabic Treebank، وبنـاء عليه تم التوصل إلى درجة خطأ كالتالي: ١٧,٣٪ على مستوى الكلمة باعتبار التشكيل الإعرابي و٧,٢٪ بتجاهله، درجة خطأ ٥,١٪ على مستوى الأحرف باعتبار التشكيل الإعرابي و٢,٢٪ بتجاهله.

واستخـدم الشافعي وآخرون [٤] نماذج ماركوف المخفية (Hidden Markov Models HMM) لحـل مسألـة التشكيـل الآلي. تطلب هذا الأسلـوب مكنزًا كبيرًا مـن النصوص المشكلة لأجـل استخراج الخصائص التـي ستستخدم في النمذجة، وقد قام بعض المؤلفين باستخـدام نصوص القرآن الكريم للنمذجة والاختبار. كما استعمـل المؤلفـون سلسلة الكلمـات غير المشكلـة كخصيصة، واعتـبروا أن الحالة المخفيـة هي السلسلة ذاتها مشكَّلةً، وبناء على ذلك؛ توصلوا إلى نسبة خطأ ٤,١٪، انخفضـت بعد ذلك إلى ٢,٥٪ باستخدام مرحلة مـا قبل المعالجة وسلاسل ثلاثية مختارة.

وقـد وصف عطيـة[٥] نظـام ArabDiac، وهو مشـكِّل تجـاري يُستخدم في النسـخ النطقي للكلام، وقد استعمـل أسلوبًا يدمج بين القواعد والإحصاءات لأجل استخـراج التشكيـل الأقـرب لنص مـا. ويعمل هذا النظـام في أربع مراحـل: تُعاد الكلمـات المختصـرة والأرقام في المرحلـة الأولى إلى ألفاظهـا، وفي المرحلة الثانية يقـوم المحلل المفرداتـي بتحديد التشكيل الأنسب للكلمة وكذلـك الأجزاء الصرفية لهـا أو تحديـد ما إذا كانـت الكلمة معرَّبـة أم لا. وفي المرحلة الثالثـة يقوم الواسم بتحديـد وسـوم أجزاء الكلام لـكل كلمة غير معرَّبة ومن ثم يقـوم المحلل الإعرابي بتحديـد التشكيل الإعرابي، أما في حالة كـون الكلمة معرَّبة فتستخدم الإحصاءات والقواعد الصوتية لاستنتاج أفضل تشكيل لها. وأخيرًا تُحوَّل النصوص المشكَّلة إلى النسخ النطقـي في المرحلة الرابعة باستخدام قواعد خاصة آخذة في الاعتبار تأثير

نطق الكلمة على الكلمة التالية لها. حصل المؤلف على دقة ٩٧٪ (لو تضع هنا نسبة الخطأ أيضا لتسهل المقارنة) دون اعتبار أواخر الكلمات، و٨٨٪ (لو تضع هنا نسبة الخطأ أيضا لتسهل المقارنة) باعتبارها.

التشكيل الآلي في اللغات الأخرى

بالرغم من أن أكثر الأبحاث المتعلقة بالتشكيل الآلي خاصة باللغة العربية؛ إلا إن عـددًا من الباحثين قد درسوا هـذه المسألة على لغات أخرى كالسندية والأردية، والتي تُشبه العربية في الكتابة، مما يجعلها مرتبطة بموضوع البحث بشكل كبير.

فقـد قام جافيد وآخـرون [٦] بدراسـة مسألة التشكيل الآلي للغـة السندية، وهـي لغة يُتحدث بها في باكستـان وأجزاء من الهند، فقامـوا ببناء نظام يستخدم الـ WordNet والتـي تخزن العلاقات الدلالية بين الكلمـات، واستخدموا فيه ثلاثة مكانز: الأول سَمَّوه CRITICAL ويستخدم للكلمات الغامضة والهامة، والثاني وقد سَمَّـوه HOMONYMY يتستخدم للكلمات الهامة المتشابهة في الكتابة المختلفة في المعنـى، أما الثالث WNL فيستخدم للكلمـات القياسية. وبحسب تجاربهم؛ حصل المؤلفـون علـى نسبة خطـأ ٧١, ٠٪ علـى مستوى الكلمـات، و٣٩, ٣٪ علـى مستوى الأحرف.

وركَّز هاريتل وآخرون [٧] على اللغات السامية وبشكل خاص اللغة السيريانية. إذ استخدمـوا طريقة نماذج ماركوف المشروطـة (Conditional Markov Models CMM) في التشكيـل الآلي والتـي تتطلب مكنزًا مشكلا، واعتمدت الكلمات (أو الأجزاء) المشكلة السابقة على الكلمة في بناء تلك النماذج. وقد حقق المؤلفون نسبة خطأ ١٥٪ على مستوى الكلمة للعربية، و٥, ١٠٪ للسيريانية.

كمـا قام علـي [٨] بدارسة المسألة للغـة الأردية، فبين كيف قـام ببناء نظام للتشكيـل الآلي قد اعتمد على مُعجم ومكنـز مشكل وموسوم يدويًا. ويمكن تلخيص

طريقة عمله كما يلي: أولًا؛ تُزال علامات التشكيل من النص قبل معالجته، ومن ثم يقوم الواسم المعتمد على HMM بوسم الكلمات، ثم يتم البحث عن الكلمة ووسمها في المعجم لكي يتم الحصول على الكلمةِ مشَكَّلةً، وإذا لم يُعثر على الكلمة يتم تحليل الكلمـة باستخدام القواعـد وإلا فتُستخدم الإحصـاءات لتشكيلها. وقد وصلت دقة هـذا النظام إلى ٩٥٪ (لو تضع هنا نسبة الخطأ أيضـا لتسهل المقارنة) بحسب ما ذكر المؤلف.

قواعد التشكيل الاستنباطية

يتميز التشكيل في اللغة العربية بدرجة كبيرة من التعقيد، إلا إن المتأمل يمكنه أن يلاحظ الكثير من القواعد والأنماط المتكررة التي تُتَّبع في التشكيل إلى حد كبير، ومن ذلك - على سبيل المثال — أن حرف التاء المربوطة ،الذي يكون دائمًا في نهاية الكلمة، لا بُدَّ أن تَسبقه حركة الفتح كما في "حَمْزَة" و"رَغْبَة"، وقس على هذا كثير من القواعد التي يمكن استنباطها عن طريق الاستقراء.

الخصائص

للتعرف علـى قواعد التشكيل الاستنباطية لا بد مـن تحديد الخصائص التي تتكـون منها هذه القواعد، فقاعدة التاء المربوطة السابقة لا تتطلب سوى معرفة أن الحرف التالي هو تاء مربوطة. ولكن يمكن للقاعدة أن تحتوي على خصائص أكثر من ذلك بكثير. لهذا قمنا بتحديد ١٦ خاصية متعلقة بالحرف الحالي وما قبله وما بعـده، والكلمة الحالية وما قبلها وما بعدها، وخصائـص أخرى كطول الكلمة وهي كالتالي:

١. الحـرف الحالي: ويمكـن أن يحتوي كل الأحرف العربيـة بجميع أشكالها وهـي ٣٦ شكلًا، وهي الأحـرف الثمان والعشرون مضافًـا إليها الهمزات بأشكالها الست، والألف المقصورة، والتاء المربوطة.

٢. موقع الحـرف الحالي: ويحسب الموقع من بداية الكلمة ويأخذ قيمًا من ١ إلى طول الكلمـة، دون أخذ حركات التشكيل والكشيدة (حرف التطويل) في الاعتبار.

٣. الحرفـان السابقان للحرف الحالي، والحرفان التاليان له: حيث يمكنهم –جميعـا– أن يكون أحد الأشـكال المذكورة في النقطـة ١، بالإضافة إلى القيمة الفارغة N إذا كان الحرف غير متوفر.

٤. حـركات التشكيل للحرفين السابقين والتاليـين (إن توفرت): وتأخذ أحد ١٤ قيمة ممكنة أو القيمة الفارغة N إن لم تتوفر.

٥. الكلمـة الحاليـة: ويمكـن أن تأخذ أحـد ١٠٠٠ قيمة هـي الكلمات الأكثر استخدامًا، وإن لم تكن ضمن تلك الكلمات فتأخذ القيمة الفارغة N.

٦. الكلمتـان السابقتان والكلمتـان التاليتان: وهي كالكلمـة الحالية في عدد القيم الممكنة.

٧. طول الكلمة الحالية (دون أخذ الحركات ولا الكشائد في الاعتبار).

استخراج القواعد

بعـد تحديد الخصائـص التي ستؤخذ في الاعتبار عند استقـراء القواعد آليًا نقـوم باستخراجها عن طريق برنامج خاص كتبناه لقـراءة نصوص المكنز المشكل سابقًا وتسجيل القواعد الممكنة دون استبعاد أيٍّ منها، ثم تخزينها في ملفات نصية تحتـوي تلك القواعد وعدد مرات تكرارها، ونسبة نجاحهـا. يوضح جدول ٢ بعضًا من هذه القواعد.

التطبيق

قمنـا باستخدام هذه القواعـد الاستنباطية لأجل تحسين نتائج التشكيل الآلي وذلك بتطبيقها على مرحلتين:

١. مرحلـة التنظيف: وهـي مرحلة استبعاد الحركات التـي يستحيل أو يندر حصولهـا. ويمكن تحديد القيمـة الأدنى التي يمكن عـن طريقها اعتبار الحركـة نـادرة أم مـن نسبة النجاح أو مـن عدد مرات التكـرار. وتكمن أهميـة هذه المرحلة في أن كثيرًا مـن أدوات التشكيل الآلي تقوم بالتشكيل بطريقـة إحصائية لا تراعي احتمالية وجود حركة ما على حرف ما. وقد لوحظ وجود الكثير من الحالات التي لا يمكن حصولها.

٢. مرحلـة التشكيل: وهـي مرحلة تطبيـق القواعد على الأحـرف التي ليس لها تشكيل، سـواء كان ذلك بسبب استبعادها في المرحلة الأولى أو بسبب نقص نسبة التشكيل في النص. ويمكن أيضًا تحديد القيمة الأدنى لنسبة التشكيل وعدد مرات التكرار التي يجب توافرها قبل تطبيق قاعدة ما.

جدول ٢ أمثلة على قواعد التشكيل الاستنباطية باعتبار ٤ خصائص هي الحرف الحالي وموقعه، والحرفان السابق واللاحق (مُرتبة بنسبة النجاح تنازليًا).

الحرف الحالي	موقع الحرف الحالي	الحرف السابق	الحرف التالي	حركة الحرف الحالي	نسبة النجاح	عدد مرات التكرار
ل	٢	ا	م	ـّ	٩٩,٧٣٥	٦٦٥٠٤٣
و	١	N	ا	ـَ	٩٩,٨٨٦	٦١٠٦٣٦
ق	١	N	ا	ـَ	٩٩,٨٥	٤٢٩٣٦٣
ل	٢	ع	ى	ـَ	٩٩,٧٧٩	٤٠٨٣٨١
ك	١	N	ا	ـَ	٩٩,٨٨	٣٠٧٩٣١
ه	٣	ن	N	ـُ	٩٩,٣٥٦	٢٧٩٢٦١
و	١	N	أ	ـَ	٩٩,٨٥٤	٢٢٩٣٧٤
ل	٢	ا	ح	ـّ	٩٩,٦٠٢	٢٢٧٩٩٢
و	١	N	إ	ـَ	٩٩,٩٢١	٢٢١٤٦١

الحرف الحالي	موقع الحرف الحالي	الحرف السابق	الحرف التالي	حركة الحرف الحالي	نسبة النجاح	عدد مرات التكرار
أ	١	N	ي	ـَ	٩٩,٩١٧	٢١٥١٦٧
و	٢	ق	ل	ـْ	٩٩,٨١٥	٢١١٣٩٨
ل	٢	ا	ع	ـْ	٩٩,٦٩٢	٢٠٥٧٤٥
و	١	N	ه	ـَ	٩٩,٤٩	٢٠٢٩٨٢
ه	٢	ل	N	ـُ	٩٩,٩٤٢	١٩٤٨٨٠
ل	٢	إ	ى	ـَ	٩٩,٨٢٥	١٩٢٧١٨
م	٢	ل	N	ـْ	٩٩,٢٥٣	١٩١١٠٢
أ	٢	ل	ن	ـَ	٩٩,٨٧٧	١٨٧٤٢٣
و	١	N	م	ـَ	٩٩,٨	١٨٣٠٤٠
ف	١	N	إ	ـَ	٩٩,٩٥٧	١٧٧٥٤٩

ويجـدر ملاحظة أننا في مرحلـة التنظيف قد استبعدنـا الخصائص المتعلقة بحـركات التشكيـل، وذلك بسـب أن النص المـراد تنظيفه قد لا يكون مُشـكَّلًا كُليًّا (علـى عكس المكنز الذي تم استخـراج القواعد منه)، مما قد يـؤدي إلى استبعاد الكثير من الحركات الممكنة بطريقة خاطئة.

طريقة التقييم

لتقييم أداء البرنامج قمنا ببناء نص اختباري تم تشكيله يدويًّا، ثم استخدمنا ٦ أدوات للتشكيـل الآلي لتشكيـل النص غير المشكّل آليًّا، وهـذه الأدوات الست هي كالتالي:

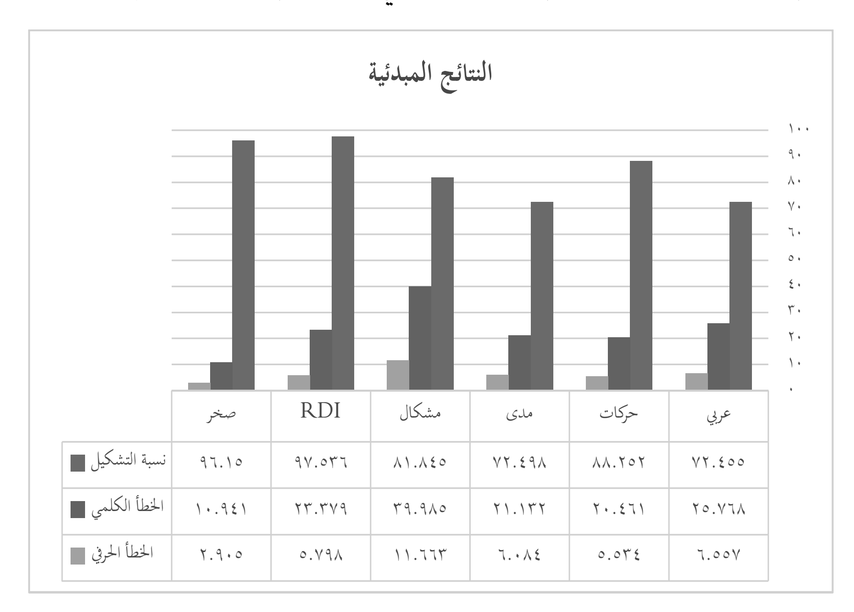
١. عربي http://www.arabinlp.com
٢. حركات http://harakat.ae
٣. مدى https://flintbox.com/public/project/8348

٤. مشكال http://tahadz.com/mishkal

٥. RDI http://www.rdi-eg.com/technologies/diac.aspx

٦. صخر http://arabdiac.sakhr.com.eg

ونظـرًا لمحدودية توافر بعض هذه الأدوات فقد قمنــا بالتواصل مع أصحابها، الذيـن وافقوا مشكورين على تشكيل النص الاختباري باستخدام أدواتهم، وإرسال النصـ المُشكَّل إلينــا. ويوضح الشكل التــالي أداء كل أداة علــى النص الاختباري. ويمكـن ملاحظة أن مُشكِّل صخر حقـق الأداء الأفضل مقارنــة بأقارنه حيث بلغت نسبة التشكيل فيه ٩٦٪ ونسبة الخطأ الكلمي ١١٪ ونسبة الخطأ الحرفي ٣٪.



	صخر	RDI	مشكال	مدى	حركات	عربي
نسبة التشكيل	٩٦.١٥	٩٧.٠٣٦	٨١.٨٤٥	٧٢.٤٩٨	٨٨.٢٥٢	٧٢.٤٥٥
الخطأ الكلمي	١٠.٩٤١	٢٣.٣٧٩	٣٩.٩٨٥	٢١.١٣٢	٢٠.٤٦١	٢٥.٧٦٨
الخطأ الحرفي	٢.٩٠٥	٥.٧٩٨	١١.٦٦٣	٦.٠٨٤	٥.٥٣٤	٦.٥٥٧

رسم توضيحي ٢ : نتائج تشكيل النص الاختباري باستخدام الأدوات الست

النص الاختباري

يتكون النص الاختباري من حوالي ٥٠٠ جملة (أو مقطع)، تم اختيارها عشوائيًا من المكنز، والذي روعي فيه التنوع ليشمل نصوصًا إخبارية ودينية وثقافية واقتصادية ورياضية. وقد تم استبعاد هذا النص من المكنز عند استخراج القواعد الاستنباطية. ويوضح الجدول التالي بعض الإحصائيات عن هذا النص.

جدول ٣ إحصائيات عن النص الاختباري

عدد الأحرف العربية	٧٧،٧٣٢
عدد الكلمات العربية	١٦،٢٤٢
عدد الكلمات الفريدة العربية	٧،٦٤٠
عدد الجمل	٤٩٥
نسبة التشكيل	٩٩,٣٪

المقاييس المستخدمة

درج الباحثون في مجال التشكيل الآلي على استخدام مقياسين رئيسين لتحديد دقة التشكيل وهما نسبة الخطأ الكَلِمي ("WER" Word Error Rate) ونسبة الخطأ الحرفي (Character Error Rate / Diacritic Error Rate)، ويتم قياسهما عن طريق المعادلتين التاليتين:

ويوضـح المثـال في الجدول التالي كيفيـة حساب المقياسـين في سياق التشكيل الآلي:

جدول ٤ مثال لتوضيح طريقة حساب نسبة الخطأ الكلمي ونسبة الخطأ الحرفي

النص الأصلي	لكل مجتهد نصيب
النص المُشكل آليًا	لِكُلِّ مجْتَهد نَصيب
التشكيل الصحيح للنص	لِكُلِّ مُجْتَهِد نَصِيبٌ
نسبة الخطأ الكلمي	66.7% ≡ 0.667 = 2 / 3
نسبة الخطأ الحرفي	17% ≡ 0.17 = 2/ 12

ورغـم أن هذين المقياسين هما الأشهر والأكـثر استخدامًا إلا إنهما لا يكفيان لمعرفة درجة أداء البرنامج الذي يقوم بالتشكيل، فمن الممكن مثلًا أن يقوم البرنامج بتشكيل نص بنسبة ٥٠٪ دون أن تزيد نسبة الخطأ في التشكيل عن ١٪. لذا رأينا أنه من الضروري استخدام مقياس جديد (بالتكامل مع المقياسين السابقين) هو نسبة التشكيل (Diacritization Level (DL)) ، والتي يمكن حسابها بالمعادلة التالية:

حيث يدخل في المُشَكَّلة الأحرف الأحرف التي لا تُشَكَّل عادة إما بسبب وضوحها كالكسرة في همزة إِنْسَانٌ، أو التي لا تَظهر نُطقًا كألف واو الجماعة وحرف الواو في جَاؤُوا.

وتجدر الإشارة إلى أن هذا المقياس يَصعُب حسابه آليًا بدقة، حيث إن الأحرف المشكلَّـة ضمنيًـا لا يمكن التعرف عليها بسهولة، لذلك فقـد قمنا باستخدام قواعد مُبسطة لتحديد تلك الأحرف، وقد حققت نتائج مقبولة إلى حد كبير.

النتائج والتحليل

لاختبـار أثر القواعد الاستنباطية ﮯ تحسين نتائـج التشكيل الآلي باستخدام الأدوات الســت سابقــة الذكر؛ قمنا بإجـراء تجربتين على النصوصـ المُشَكَّلة آليًّا: الأولى باستخدام حد أدنى للندرة ١٪، والثانية باستخدام حد أدنى ٠٪. ويُمثل هذا الحد الأدنى الدرجة التي يتم عندها تحديد ما إذا كان تشكيل حرف معين نادرًا أم لا؛ حيث إنه كان نادرًا فإنه يتم استبعاده. ويُظهر الرسمان التوضيحيان ٣ و٤ نتائج كلتـا التجربتين باستخدام المقاييس الثلاثة: درجة التشكيل، ونسبة الخطأ الكَلِمي، ونسبة الخطأ الحرﮯ.

ويمكننا أن نُلاحظ أن نسبة التشكيل تتحسن بشكل كبير (ﮯ التجربتين) على أغلـب الحالات خاصــة أن أكثر هــذه الأدوات لم تقم بتشكيل النصـ بشكل كامل، وذلـك باستثناء مُشكِّليّ صخر وRDI الآليّن اللذين انخفضت فيهما نسبة التشكيل بشكل طفيف.

أمــا بالنسبــة لنسبة الخطأ الكلمــي فنرى انخفاضًا يــتراوح بين ١-٢٪ لأغلب الحــالات، وقـد يصل إلى ٦٪ ﮯ حالة أداة مشكال، وزيــادة ١٪ تقريبًا ﮯ حالة أداة صخــر. أما ﮯ التجربة الثانيــة، حيث قمنا بخفض الحد الأدنــى للندرة، فنلاحظ انخفاضًا ﮯ حدود ١٪ لجميع الأدوات الست.

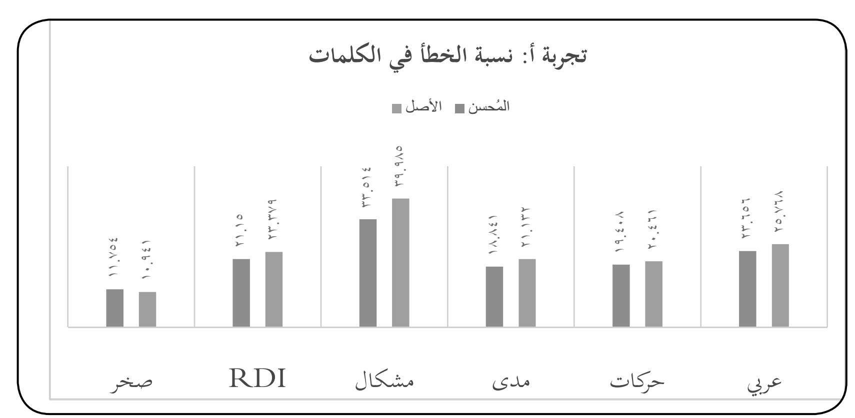
وأخـيرًا، نــرى نسبة الخطــأ الحــرﮯ تنخفض أيضًــا بحــوالي ٠,٥٪ ﮯ كلتا التجربتين، لأغلب الحالات باستثناء حالة مشكّل صخر.

ويُلاحـظ من هذه النتائــج أن انخفاض نسبة النُدرة يزيـد من نسبة التشكيل بشكل كبير، حيث وصلت إلى حوالي ٩٥٪ ﮯ جميع الأدوات الست، وهذا يُمثل تغيرًا كبــيرا يصــل إلى أكثر من ١٠٪ لجميع الأدوات باستثنــاء صخر وRDI. ويعود ذلك

لكــون خفض نسبة الندرة يُبقي على درجة أكبر من التشكيل الأصلي، وذلك قبل أن يتم تطبيق القواعد الاستنباطية على النص.

ويُلاحــظ كذلك أن انخفاض نسبة النُدرة يُقلِّــل من التحسن في نسبة الخطأ، سواء الكلمي أو الحرفي، حيث يؤدي ذلك إلى الإبقاء على أخطاء التشكيل الموجودة مُسبقًا دون إزالتها.

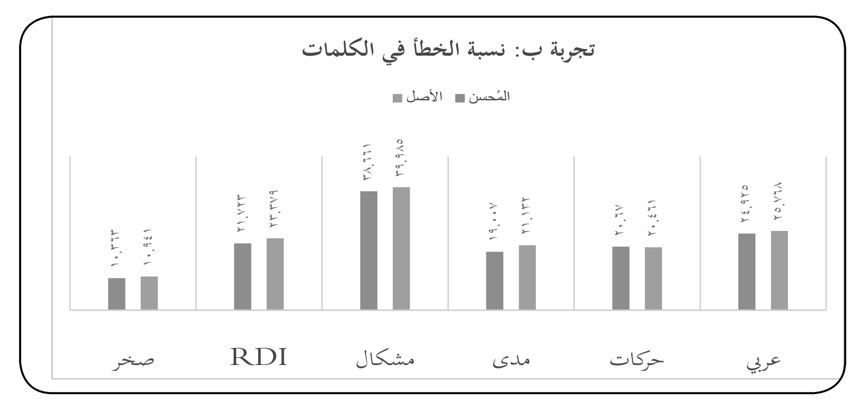


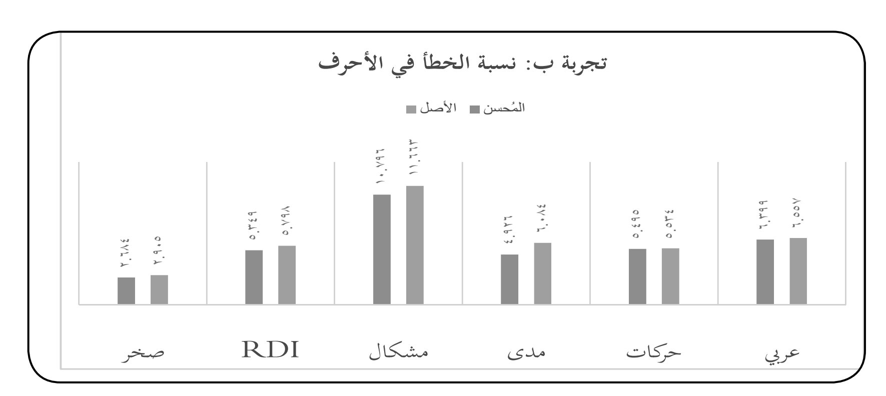




رسم توضيحي ٣: نتائج استخدام القواعد الاستنباطية في التجربة أ (حيث تم استخدام ١٪ كحد أدنى في مرحلة التنظيف)







رسم توضيحي ٤ : نتائج استخدام القواعد الاستنباطية في التجربة ب (حيث تم استخدام ٠٪ كحد أدنى في مرحلة التنظيف)

ويُمكننـــا أن نقـــارن درجة التحســـن لِكلّ مقياس من جـــدول ٥ وجدول ٦. حيث فمنا في حالة المقياس الأول (نسبة التشكيل)؛ بحساب درجة التحسن بطرح الفرق بين القيمتين ثم قسمته على القيمة الأصلية. أما في حالة نسبة الخطأ (سواء كانت في الكلمـات أو الأحرف) فقد قمنا بضرب الناتج بإشـارة سالبة، لأن التحسن هنا هو درجة النقص وليس درجة الزيادة.

جدول ٥ درجة التحسن في التجربة أ

اسم الأداة	نسبة التشكيل	نسبة الخطأ (كلمات)	نسبة الخطأ (أحرف)
عربي	٢٧,٨٪	٨,٢٪	٧,٢٪
حركات	٥,٩٪	٥,١٪	٨,٤٪
مدى	٢٧,٨٪	١٠,٨٪	١٩,٧٪
مشكال	١١,٤٪	١٦,٢٪	٢٣,٠٪
RDI	-٣,٥٪	٩,٥٪	٨,٧٪
صخر	-٣,٢٪	-٧,٤٪	-٥,٠٪
المتوسط	١١,٠٣٪	٧,٠٧٪	١٠,٣٣٪

جدول ٦ درجة التحسن في التجربة ب

اسم الأداة	نسبة التشكيل	نسبة الخطأ (كلمات)	نسبة الخطأ (أحرف)
عربي	٣٢,٢٪	٣,٣٪	٢,٤٪
حركات	١٠,٦٪	-١,٠٪	٠,٧٪
مدى	٣٢,٣٪	١٠,١٪	١٩,٠٪
مشكال	١٥,٤٪	٣,٣٪	٧,٤٪
RDI	-٠,٢٪	٧,١٪	٧,٧٪
صخر	-٠,٥٪	٥,٣٪	٧,٦٪
المتوسط	١٤,٩٧٪	٤,٦٨٪	٧,٤٧٪

الخلاصة

في هـذا العمل البحثي قمنا بتقديم طريقـة القواعد الاستنباطية، وهي قواعد تشكيليـة تم استخراجها من أحد المكانز، وتحدد مـدى تأثير عدد من الخصائص (كنـوع الحرف وموقعه في الكلمة) على التشكيـل. قمنا باستخدام تلك القواعد في مرحلتـين: مرحلـة التنظيف وهـي المرحلة التي يتـم فيها استبعـاد التشكيل حسب درجـة نُدرتـه، ومرحلة التطبيق وفيهـا يتم تطبيق هذه القواعـد على الأحرف غير المُشكَّلة في النص.

وتظهـر نتائج التجارب التي قمنا بها تحسنًا كبيرًا في نسبة التشكيل تصل إلى ١١٪ في المتوسـط، وكذلك تحسنًـا في نسبة الخطأ الكلمـي بمتوسط ٧٪، وفي نسبة الخطأ الحرفي بمتوسط ١٠٪.

المراجع

1. M. Rashwan، M. Al-Badrashiny، M. Attia، S. Abdou and A. Rafea، "A Stochastic Arabic Diacritizer Based on a Hybrid of Factorized and Unfactorized Textual Features،" Audio، Speech، and Language Processing، IEEE Transactions on، vol. 19، no. 1، pp. 166-175، jan. 2011.

2. Wikipedia، "Diacritic --- Wikipedia، The Free Encyclopedia،" 16 January 2012. [Online]. Available: http://en.wikipedia.org/wiki/Diacritic. [Accessed 1 February 2012.[

3. I. Zitouni and R. Sarikaya، "Arabic Diacritic Restoration Approach Based on Maximum Entropy Models،" Computer Speech & Language، vol. 23، no. 3، pp. 257-276، 2009.

4. M. Elshafei، H. Al-Muhtaseb and M. Al-Ghamdi، "Statistical Methods for Automatic Diacritization of Arabic Text،" 2006.

5. M. Attia، "Theory and Implementation of a Large-Scale Arabic Phonetic Transcriptor، and Applications،" PhD. Dissertation، RDI، Cairo، Egypt، 2005.

6. J. Mahar and G. Memon، "Lexicon Based Diacritic Resto-

rations Using WordNet For Sindhi،" International Journal Of Academic Research، vol. 3، no. 2، pp. 37-43، March 2011.

7. R. A. Haertel، P. McClanahan and E. K. Ringger، "Automatic Diacritization for Low-Resource Languages Using a Hybrid Word and Consonant CMM،" Stroudsburg، PA، USA، 2010.

8. A. Ali، "Automatic Urdu Diacritization،" M.Sc. Thesis، Center for Research in Urdu Language Processing، Lahore، Pakistan، 2009.

9. O. Shaaban، Automatic Diacritics Restoration for Arabic، Dhahran، King Fahd University of Petroluem and Minerals، 2013.

الترجمة الآلية من العربية وإليها

أ.د. محمد زكي خضر (*)

khedher@ju.edu.jo

(*) الجامعة الأردنية

حصـل علـى البكلوريوس في الهندسـة من جامعة بغداد عـام ١٩٦٥ وعلى الدكتـوراه من بريطانيا عام ١٩٧٢ وعمل في جامعة الموصـل بالعراق وحصـل على مرتبة الأسـتاذية عام ١٩٨١، وهو يعمل في الجامعـة الأردنية منذ عام ١٩٩٢. نشـر أكثر مـن ١٠٠ بحث وألف ٢٠ كتابا. عمل في مجال المعالجة الآلية للغة العربية ونال جوائز عالمية وهو عضو في جمعيات علمية عالمية ويرأس تحرير المجلتين الدوليتين للتطبيقات الإسلامية في علم الحاسوب وتقنياته باللغتين العربية والإنجليزية.

ملخص

حصل تقـدم هائل للترجمة الآلية خلال السنـوات الماضية خاصة بين اللغات الأوربيـة. وفي الوقـت نفسه حصـل تقدم كبـير في الترجمة الآلية للغـات الشرقية كالصينيـة واليابانيـة والكوريـة. كمـا توفـرت أنظمـة متعـددة مساعـدة لأعمال المترجمـين المختصـين وأنظمـة التصحيح اللغـوي وغيرها، مما يساعـد في عملية الترجمة اليدويـة. أما الترجمة من اللغة العربية وإليهـا فهناك محاولات للترجمة الآلية، لكن معظم هذه الأنظمة تنتج لغة ركيكة وغير دقيقة.

تعتمـد أنظمة الترجمة الآلية علـى أسس مختلفة، فبعضها يستند إلى المعاجم مع مساعدة نحوية لتحليل الجملة ثم إعادة تركيب ترجمتها وبعضها يعتمد الترجمة بالاستنـاد إلى الذخيرة اللغويـة المحتوية على نصوص ترجمـت من قبل مترجمين من البشـر باستعمال عمليات إحصائية وغير ذلك من الوسائل حيث تمثل المعاجم ركيزةً رئيسيةً في الترجمة، إضافة إلى البرمجيات اللغوية المتخصصة.

هـذا البحث هو استكمال لبحث سابق (١) ويهدف إلى تقديم فكرة عن تأريخ الترجمة الآلية بين اللغات العالمية وتطورها وما وصل إليه التقدم في هذا المجال كما يناقشـ الوضع بالنسبة للترجمة الآلية من اللغة العربية وإليها والمشاكل التي تعاني منها ومن ثم يبين الخطوات التي ينبغي اتباعها للتقدم في هذا المجال، ثم يعرج إلى مؤشـرات عن مستقبل الترجمة الآلية من اللغـة العربية وإليها وما يحتاج ذلك من جهود.

مقدمة

إن اختـلاف اللغـات التي يتكلم بهـا البشر آية من آيـات الله تعالى في الأرض قـال تعالى: "وَمِنْ ءَايَاتِهِ خَلْقُ السَّمَاوَاتِ وَالأَرْضِ وَاخْتِلاَفُ أَلْسِنَتِكُمْ وَأَلْوَانِكُمْ إِنَّ فِي ذَلِكَ لأَيَاتٍ لِّلْعَالِمِينَ" (الروم/٢٢)، ولذلك احتاج الناس الترجمة منذ الزمن الغابر الـذي تعددت فيه لغات البشر لغرض التواصل بينهم، فعن خارجة بن زيد بن ثابت قال: قال زيد بن ثابت: "أمرني رسول الله -صلى الله عليه وسلم- فتعلمت له كتاب يهود، وقال: إني والله ما آمن يهود على كتابي، فتعلمته، فلم يمر بي إلاّ نصف شهر حتى حذقته، فكنت أكتب له إذا كتب، وأقرأ له إذا كُتِب إليه" -رواه أبو داود-.

الترجمـة هـي نقـل معاني نصـ من لغـة إلى لغـة أخرى مـع مراعـاة للدقة والأسلـوب. ويحتاج ذلك فهـم النص الأصلي والتعبير عن المحتـوى والأسلوب بلغة أخرى. فالمترجم يجب أن يتقن اللغتين المترجم منها والمترجم إليها وهو يتبع إحدى طريقتـين: إما الترجمـة الحرفية والالتزام بمعاني مفردات النصـ الأصلي ونقلها إلى اللغـة الثانيـة أو فهم المعنى العام ثم التعبير عنه باللغة الثانية بأسلوب المترجم نفسه (١).

يتعـرف المترجم على الرموز المكتوبة في الترجمة الكتابية والأصوات المنطوقة في الترجمـة الشفويـة، لكن اللغة العربيـة المكتوبة بحروف غـير مشكولة تستوجب معرفـة من القـارئ لكي يفهم المقصود بـدون تشكيل. وهو بذلـك يستعمل خبرات ذاتية من قواعد اللغة لكي يستعملها بشكل ذاتي وسريع. أما المترجم للكلام المنطوق فهـو يستطيع تمييز الجملـة التي يترجمها إن كانت استفهامية أم خبرية أم تعجبية دون وجود علامات استفهام أو تعجب (٢).

بعد ذلك يرجـع المترجم ذهنيًا أو واقعيًا إلى الوحدات المعجمية وهي الكلمات والتعبيرات الاصطلاحية ويفهم معانيها في سياقاتها اللغوية والاجتماعية المختلفة،

فالكلمـة الواحدة قد تعني أشياء كثيرة يقيدها السياق الذي ترد فيه. فمثلًا "عين" قـد تعني عـين الإنسان أو الحيوان وقـد تعني عين الماء وقد تعنـي في الصرف عين الفعـل (٢). وهذه الترادفات تمثل مشكلة كبـيرة بالنسبة للترجمة عامة والترجمة الآلية خاصة.

ولحـل هذه المشكلة يقوم المترجم بتحليل المفهـوم للوصول إلى كنهه وما يتفرع عنه من ظلال في المعنى، وخاصة المجازية منها ، وذلك حتى تتحدد العلاقة بين هذا المفهوم الأساس وما يتفرع منه وبين المفاهيم التي يمكن أن تتصل به بشكل أو بآخر، أي العلاقة بينه وبين مجموعات المفاهيم الأخرى.

فالمفهوم الذي يرتبط بجسم واحد بعينه، أو يدل عليه، هو مفهوم إفرادي، مثل "عطارد". أما إذا ارتبط المفهوم بعدة أجسام بينها نوع من التجانس الذي يضعها في مجموعة واحـدة، فإن المفهوم يكون عامًا (بمعناها الضيق هنا)، مثل "كوكب". فهذا المفهوم يدل على عدد من الأجسام التي تدور حول الشمس. ولتحديد المقصود يمكـن أن يكـون التعريـف مركـّزًا أو موسعًا. وعلى ذلـك فإن دقـة التعريف تحدد المفهـوم. والتعريـف قد يكون شامـلًا في بعض الأحيان، وفي غالـب الأحيان لا يخلو مـن النقص أو أن يكون تعريفًا بالضـد أو ما يعرف بالتعريف السلبـي أو أن يكون فضفاضًـا أو ضيقًـا أو يأخذ بمبدأ الإحلال والتعويض إلى غير ذلك من الصفات التـي يتصـف بها التعريف. وهـذا التعريف قد يكون واضحًـا في ذهن المترجم وقد يكـون مشوشًـا فيتصرف في ترجمته لهذا المفهوم وقد يعـبر عنه بمفهوم يفهمه هو ويكون قريبًا أو بعيدًا من المفهوم الأصلي.

ولقـد سبق الإمام أبو حامد الغزالي اللسانيين المحدثين في تحديد العلاقة بين المفهـوم واللفظ حيث يقـول: "الشيء له في الوجود أربـع مراتب: الأولى حقيقته في نفسـه والثانيـة ثبوت مثـال حقيقته في الذهن وهـو الذي يعبر عنـه بالعلم والثالثة تأليـف صوت بحـروف تدل عليه وهـو العبارة الدالـة على المثال الـذي في النفس،

والرابعة تأليف رقوم تدرك بحاسة البصر دالة على اللفظ وهو الكتابة، فالكتابة تبع للفظ إذ تدل عليه، واللفظ تبع للعلم إذ يدل عليه، والعلم تبع للمعلوم إذ يطابقه ويوافقه، وهذه الأربعة متطابقة متوازية، إلاّ أن للأولين وجودان حقيقيان لا يختلفان بالأعصار والأمم، والآخرين وهما اللفظ والكتابة يختلفان بالأعصار والأمم لأنهما موضوعان بالاختيار" أ.ه (٣)

كان للتطور في تمثيل اللغويات الذي قاده نعوم شومسكي تأثيرًا على نظريات اللسانيات الحديثة. فقد أعاد شومسكي تعريف الأهداف من النظريات اللسانية لكي تأخذ بعين الاعتبار خلفية المتكلمين بالنسبة للغتهم بدل البحث في المكانز اللغوية للوصول إلى المطابقة المطلوبة. فالفرد قد يسمع جملة بلغته الذاتية فيفهم معناها بسهولة رغم أنه لم يسمعها من قبل (١٤).

إن قواعد أية لغة محدودة لكن الجمل التي يمكن أن تركب باستخدام تلك القواعد لا نهاية لها. أما المتكلمون باللغة فيتعلمون بالاستنتاج، فمثلا يمكن للمتكلم أن يشعر بالفرق بين معنيين مختلفين في جملة غامضة مثل "حضر ابن عمي الصغير" فيمكن أن يقصد أن الذي حضر هو الابن الصغير للعم أو هو ابن أصغر الأعمام. إن القواعد اللغوية يجب أن تحاكي قابلية المتكلم بلغته الأم بحيث تستطيع أن تميز الغموض في مثل هذه الجملة وأن تعطي نموذجين للتركيبتين المحتملتين.

وقد أجبرت الحواسيب اللسانيين أن يكونوا أكثر دقة في وصفهم لقواعد اللغة التي سبق وأن تصوروا أنهم يعرفونها عن اللغة. ولكن لا يزال الشوط بعيدًا في دقة القواعد التي وصلوا إليها لتمثيل اللغة بشكل صادق.

إن التقدم في الوصف الدقيق والمعمق للغات الطبيعية وتوفر التقنية الحاسوبية لمعالجة اللغات الطبيعية اجتمعا في تكوين بيئة مكنت من الوصف المعمق للغات مع أسس قواعدية لتلك اللغات.

تساعد بعض العمليات في التعرف على المفهوم مثل التعرف على الوحدات النحوية وعلى وظائفها مثل المسند والمسند إليه والتكملة والمضاف والمضاف إليه، والجار والمجرور والصفة والموصوف وغير ذلك. وعلى المترجم أن يفهم معنى كل واحد من هذه العناصر ووظيفته اللغوية والنحوية في النص. فالصيغة الخبرية مثلا قد تأتي للإخبار كما تستعمل للطلب غير المباشر كما في عبارة: "أرى أن لديك قلمًا زائدًا"، بل وللدعاء كذلك مثل: "رحم الله فلانًا" (٤). ومن الأمور المهمة في الترجمة معرفة حقل النص ومجال تخصصه. مثلًا إذا كان النص في علم الأحياء فإنك تجد أن بعض الكلمات تختلف معانيها في هذا النص عن حقل تجارة المواد الغذائية أو في حقل من حقول العلوم الزراعية. وهكذا.

المحور الثاني في الترجمة هو التعبير عن محتوى النص باللغة الهدف أي اللغة المترجم إليها. ويحتاج هذا معرفة بإنتاج المقابلات الصوتية أو الكتابية والنحوية والمعجمية والبلاغية. وعلى المترجم أن يقوم باختيار العبارات المناسبة في اللغة التي يترجم إليها التي توافق المفهوم الذي توصل إليه في الخطوات السابقة. "وعند ذلك تبرز براعة المترجم في اللغتين حيث يجب أن يتقن ما يسمى بالتقابل المعجمي أو الاصطلاحي، فالتعبير العربي (رجع بخفي حنين) ليس له مقابل حرفي بأية لغة، ولو ترجم ترجمة حرفية لأصبحت الترجمة مضحكة. فهو يجب أن يفهم مثل هذه العبارات ويجب أن يعرف العبارات الدارجة التي تقابلها في اللغة الثانية أو على الأقل يكوّن هو تعبيرًا من كلمات تلك اللغة بما يعبر عن المفهوم بدقة كافية مستعملًا القواعد النحوية والصرفية في اللغة التي يترجم إليها" أ.ه (١).

أما التعامل مع الجمل الطويلة فهو من أهم مشاكل الترجمة حيث يحتاج ذلك إلى معالجات ذهنية معقدة. فحينما تحوي الجملة الطويلة على عدة مفاهيم فإن المفهوم الواحد ربما يتجزأ ليتكامل بين بدء الجملة ونهايتها. وقد يتضمن المفهوم الواحد مفاهيم ثانوية. وقد يكون المفهوم المهم مندسًا بين عدد من المفاهيم الأقل

أهميــة منه. وهكذا فإن الجملة الطويلة ربما تستعصي على الفهم والتحليل وتحتاج براعة عند التركيب في اللغة المترجم إليها.

يمكننــا أن نذكـر أن المترجمــين لا يقومون كلهم بالخطـوات نفسها، بل إن كل مترجم له أسلوبه الذي يتبعه في الترجمة. ماذا نفعل تمامًا حينما نحاول فهم نص ما من النصوص؟ إن الطريقة التي يتعامل بها الإنسان ذهنيًا مع النصوص ما زالت غـير معروفة على وجه الدقـة وتخضع لنظريات غير مؤكدة حتى الآن. وليس هناك أساليــب متفق عليها حتى الآن يمكن أن ننقلها بشكل من الأشكال إلى الحاسوب في الترجمــة الآلية. لذلك فإن فهم النص يعتــبر مشكلة كبرى في الترجمة، فهناك في كل اللغات لبس دلالي نتيجة تعدد معاني بعض الكلمات وتحديد الجهة التي تعود لها بعض الضمائر (١). وهناك غموض آخر بسبب بناء أو تركيب الجملة يسمى اللبس النحوي، حيث تعني الجملة أكثر من معنى أو يمكن تفسيرها بأكثر من طريقة، كما أن هنــاك لبس على المستـوى الصرفي خاصة في الأفعال المتعديــة لمفعولين والصيغ التي تعني اسم الفاعل واسم المفعول في الوقت نفسه وصيغ المبني للمجهول والربط بين عدة وحدات لغوية مثل تغير معنى الفعل حسب حرف الجر الذي يأتي بعده مثل "رغب في" و"رغب عن"، وهناك مصادر أخرى عديدة للبس (٥).

إن حـرف الجر قـد يأخذ معاني مختلفة بين اللغـات المختلفة، فمثلا اللام في جملة: "أعطيت تفاحة لأحمد" تعني اللام "إلى=to" بينما في جملة "أكلت تفاحة لأحمد" تعني أن التفاحة هي "ملك أحمد=of".

الحاسوب في خدمة الترجمة

الترجمة الآلية اليوم لا ترقى إلى المستوى الذي يضاهي المترجمين من البشر. لكـن الحاسوب اليوم بإمكانه أن يقدم عونًا لا بأس به للمترجم. ومن هذا العون ما يأتي: (١)

١. الترجمـة الآليـة التي تحتـاج إلى تحرير لاحق، أي الحاجـة إلى مراجعة بشرية بعد الترجمة الآلية.

٢. تحريـر أولي ثـم القيـام بالترجمـة الآلية بعـد التحرير السابـق، وذلك بتبسيط الجمل الطويلة وتحديد معاني الكلمات ذات المعاني المتعددة.

٣. التحـاور مع الآلة بـين المترجم والحاسوب، حيث هنـاك برامج حاسوبية ذوات إمكانيـات حوارية بأن تعطي الترجمـة جملة بعد جملة، ويتوقع من المترجم أن يوافق أو يعدل عليها.

٤. قيام المترجم بالترجمة وتساعده الآلة، كأن تعطي الآلة للمترجم الكلمات ذات العلاقة من المعجم مع مرادفاتها، وهو يختار من بينها.

٥. الخدمات الحاسوبية الأخرى وسنورد بعض ما يتوفر من برمجيات تخدم الترجمة في ما بعد.

نشأة الترجمة الآلية وتطورها

كان أول مـن استخـدم الحاسـوب في الترجمـة فعليًـا هـو وارن ويفـر عـام ١٩٤٧، فقـد كتـب في عام ١٩٤٩ مذكـرة لزملائه بأربعة مقترحـات لتطوير أنظمة الترجمـة الآلية لكـي تتقدم أكثر مـن الترجمة كلمـة كلمة وذلك بفحصـ المفهوم المباشـر للكلمات الغامضة. كما أشار إلى التناظـر بين تركيب المخ البشري والآلات المنطقية ثم انتهى بأن الترجمة الآلية ممكنة التحقيق. وفي تلك الفترة كانت هناك حاجـة سريعة للترجمة غـير الدقيقة نظرًا لحاجة الولايـات المتحدة للاطلاع على اتصـالات الاتحاد السوفياتـي، وفي الوقت نفسه ظهرت الحواسيـب. ولذلك كانت أنظمـة الترجمـة من الروسيـة إلى الإنكليزية أولى المحاولات، فقـد أجريت بنجاح في جـورج تـاون أول ترجمة من اللغـة الروسية إلى الإنكليزية عـام ١٩٥٤، وفي عام ١٩٥٥ أجريت في الاتحاد السوفياتي أول تجربة في الترجمة الآلية من الإنكليزية إلى

الروسية باستعمـال قاموس يحوي ٢٣٠٠ كلمة. وأشارت التجربة إلى ضرورة تمثيل التراكيـب اللغوية بمستوى بناء الجملـة وبالمستوى المعجمي، كمـا اتضح أن مشكلة الغموض هي مشكلة فعلية رغـم أنها لم تقدر حق قدرها حينئذ، حيث كانت الفترة مـن ١٩٥٦ إلى ١٩٦٦ فترة التوقعات الكبيرة من الترجمـة الآلية التي لم يتحقق ما كان يؤمـل مـن الحصول على دقة بنسبة ٩٥٪ دون معالجـة مسبقة للنصوص. لقد أوضحـت الأبحاث أن تركيـب اللغات أكثر تعقيدًا مما كان متصورًا، وأن المترجمين مـن البشر يستعملون الكثير مـن المعرفة في حقل المادة المترجمـة والمعرفة بالحياة العاديـة أكثر ممـا كان متصورًا. وأدى ذلـك إلى تباطؤ في أبحـاث الترجمة الآلية وربما إهمال الموضوع حتى عام ١٩٧٥ (١).

عادت الأبحـاث على أشدها في أوربا وكندا واليابـان بين عامي ١٩٧٥ و١٩٨٥ باستعمـال الأنظمة الخبيرة وأبحاث معالجة اللغات الطبيعية، وبشكل خاص للغات الأوربية واليابانية، وبذلك ظهر الجيل الثاني من برامج الترجمة الآلية فظهرت في الأسـواق برامج تجاريـة تستعمل الحواسيب. وهذا التطور شمـل البحوث المعجمية والنحو والصرف والدلالة.

وفي عام ١٩٨٩ كانت بداية الجيل الثالث من برامج الترجمة الآلية التي استندت إلى المعلومـات الإحصائيـة حينمـا قامـت شركـة IBM (٢٩)، بمشـروع Candide المستنـد إلى الترجمة بالأمثلة والترجمة المحدودة الموضـوع مع تعدد اللغات المترجم منهـا والمترجم إليها، فـازداد الاهتمام بالترجمة الآلية لحـد لم يسبق له مثيل. وقد استنـدت هـذه الترجمة إلى الذخيرة اللغوية Corpus Based والتي لا تزال في تطور حتى اليوم مع بعض التكامل مع الأساليب الأخرى للترجمة الألية (١).

أمـا الترجمة مـن اللغة الإنكليزية إلى العربية فقـد كان أول ظهورها في نهاية السبعينيـات في ولاية يوتـا الأمريكية، ولم يكن يحتوي البرنـامـج على تحرير أولي، لكـن كان بالإمكان إجراء تحريـر نهائي. وقد احتوى النظـام على مرحلتين الأولى لتحليـل اللغـة المصـدر والثانية لتوليـد اللغة الهـدف. كان تحليل اللغـة الإنكليزية

موجهًـا لكي يولـد العبارات في اللغة الهـدف بشكل مباشر باستعمـال معجم ثنائي اللغة للعبـارات. ولم تكن الألفاظ والجمل الإنكليزية تحلل بعمق، بل فقط بما يكفي لتوليـد المقابل باللغة العربية. وهكذا كان النظام باتجاه واحد ولم يكن يغوص بعمق في اللغة المصـدر. وقد استعمل البرنامج لترجمة الموسوعة البريطانية (بريتانيكا) إلى اللغـة العربية (١٤). كما قامت سلطنة عمـان بالحصول على امتياز للبرنامج لكي يستعمل في ترجمة الوثائق الإنكليزية إلى العربية.

لم تكـن الطريقة المباشـرة في الترجمة تغوص في أعماق اللغـة المصدر، ومن ثـم أدرك مطورو البرمجيات أن الطريقة المباشرة لا تكفي لمعالجة اللغات الطبيعية المعقـدة. فقد وجـد أن التعامل مـع الجمل باستبـدال مواقع الفعل بـين الإنكليزية والعربيـة مثل Ali drinks coffee فترجمتها إلى: "يشرب علي القهوة" احتوى على استبدال موقع الفعل والفاعل لكن ذلك لا يكفي في جملة مثل:

The boy whose clothes are dark and carrying a flag in his right hand talked to me.

فمثل هذه الجملة تحتاج إلى تحليل عميق لتحديد الفاعل ومن ثم اقترانه بالفعل حيث بينهما عدد كبير من الكلمات. وعلى ذلك فإن الطريقة المباشرة لا تعطي نتائج دقيقـة لمثل هذه الجمـل الطويلة، لذا فقد برزت حاجة لتطويـر المعرفة بكيفية تحليل الجمل الطويلة بعمق وكفاءة دون غموض، وعند ذلك ظهرت الطريقة التحويلية التي قدمت للترجمة الآلية أمرين: الوصف لتحليل الجملة ومن ثم التقنية الجديدة لتمثيل هذا التحليل العميق، وهو ما طور خلال السبعينيات والثمانينيات من القرن الماضي.

إن استعمـال أنظمـة الترجمة الآليـة قليلة الدقـة يدفع إلى -بشـكل مطرد- للاستعانـة بمترجمين من البشـر. وقد وظفت بلدان كثـيرة الترجمة الآلية لخدمة متطلباتها الاجتماعية والاقتصاديـة والعلمية والتقنية في طليعتها الولايات المتحدة وروسيا والصين واليابان (١و٢٣).

لقـد أجريـت منـذ الثمانينيـات في القرن الماضـي العديد مـن الأبحاث حول المعالجـة الصرفيـة للغة العربيـة وكان معظم تلـك الدراسات ينصـب على عملية فصـل السوابـق واللواحـق في الكلمة العربيـة المركبة، ومن ثـم استخلاص الجذر لمعرفـة العمليات الصرفية التي أجريت عليها والتي يمكن للحاسوب القيام بها بعد برمجته بشكل دقيق.

قام كيفن نايت ودانيال ماركو بتحويل الأبحاث العلمية المتعلقة بالترجمة الآلية إحصائيًـا في عام ٢٠٠٢ إلى منتـج للترجمة الآلية من العربية إلى الإنكليزية، وكان ذلـك مثالاً جيدًا لتكويـن ذخيرة لغوية متقابلة بين العربيـة والإنكليزية والاستفادة منها للترجمة الآلية.

كما وجـد أن استكشـاف أسماء الأعـلام والبيانات المعروفـة ببرنامج خاص يمكـن أن يحسـن من الترجمة، فمـا أن يكتشف مثل هذا الاسـم حتى يمكن القيام بعدة عمليات تفيد في الترجمة، فمثلا عند اكتشاف أن كلمة مثل "أحمد" هي اسم علم فإنها لا تترجم على أنها I praise

كمـا ظهـرت بعض الأنظمـة الهجينية (٣٤) التـي تستعمل للترجمـة المكتوبة والمنطوقـة باستعمـال الطريقتـين: الطريقة الإحصائيـة والطريقـة المستندة إلى القواعـد (٢٤و٢٥)، وبعد النظر إلى كل من الطريقتين من ناحيتي الفوائد والمضار يمكن المزاوجة بينهما للحصول على الفوائد من كليهما واستبعاد المضار (١٢).

هناك الآن في الأسواق أعداد كبيرة من برامج الترجمة الآلية بين اللغات الأوربية، وقـد ازداد الطلـب عليها بشكل واسـع وساهمت شبكة الإنترنيت بزيـادة الحاجة إلى الترجمـة الآليـة وسهلت في تبادل المـادة المترجمة إلى من يحتاجهـا. ولا يزال التقدم على أشده في تكامل عمل المترجمين من البشر مع الترجمة الآلية، حيث يحتاج التقدم بمستوى الترجمة الآلية إلى مترجمين أكفاء لتطويرها والارتقاء بها (٧).

مزايا الترجمة الآلية

يحتاج المجتمع البشري اليوم إلى كمية هائلة من المعلومات التي يراد ترجمتها بمـا يفوق ما يتوفر من مترجمين من البشر خاصــة بعد تفجر ثورة المعلومات وتنوع اللغـات التي تساهم في إنتاج المعرفة اليـوم وما يحتاجه من لا يتكلمون تلك اللغات. ورغـم التأريخ الحافـل للغة العربية في مساهمتها في المعرفـة البشرية، لكنها اليوم ليســت المصدر الأهــم في العلوم والتقنيـات الحديثة. وهذا يجعـل عملية الترجمة مــن اللغات الأخرى إلى اللغة العربيــة ذات أهمية بالغة للناطقين بالعربية، فإمّا أن يتعلمـوا لغة أو لغات أخـرى بجانب العربية أو أن يترجم لهــم ما يصدر من معرفة باللغـات الأخرى. فالتخلـف في الترجمة إلى اللغــة العربية يفســر الإقبال الشديد علــى تعلم اللغة الإنكليزية خاصة والتدريس بها في الجامعات والمدارس في كثير من الأقطار العربية (١).

يضــاف إلى ما سبق سبب ثــان في أهمية الترجمة الآلية هو أن عملية الترجمة مملــة للمترجمــين وتستهلك الأوقات الطويلة منهم، فهــم يحاولون التغلب على الملل بتغيير أسلوب الترجمة أو بالراحة. والمترجم حاله حال البشر، ينام ويمرض ويغير عملــه ويتقاعــد، وهذا يجعل المترجمــين من البشــر بضاعة نادرة. كمــا أن غالبية المترجمين يتقنون لغة واحدة مع اللغة الأم. وهذا بخلاف الآلة التي يمكنها أن تعمل الوقــت الطويل ويمكن استبدالها بما هو أكفأ منها مع التقدم التقني المستمر إذا ما تقادمت (١).

وهنــاك أسبــاب أخــرى منهــا أن المؤسســات الضخمــة ترغــب في استعمــال مصطلحــات وعبارات محددة ينبغــي أن تعاد كلما مرت تلك العبارات، والمترجمون مــن البشر عــادة يحاولون تغيــير أساليبهم ابتعادًا عــن الملل، إضافــة إلى أسباب اقتصاديــة وتجارية في استعمال الترجمة المستنــدة للحاسوب حيث تزيد من حجم ومن سرعة ما يراد ترجمته مما يحقق للمؤسسات والشركات إنتاجية عالية بسرعة هائلــة، كما أن حاجة القطــاع التجاري قد تكون إلى ترجمــة تعطي فكرة عامة عن

المنتجـات الصناعية والتجاريـة دون أن تكون على درجة عالية من الرصانة. يجري هـذا مع ازدياد حجم مثل هـذه المعلومات المطلوب ترجمتهـا وتوسعها لتشمل لغات جديـدة في عصـر العولمة. وهـذا ما يجعل الترجمـة الآلية رغـم انخفاض كفاءتها مفيدة (١و٢٢).

٢٥. مستويات برامج الترجمة الآلية وأنواع منها

يمكـن تصنيـف برامج الترجمـة الآلية بحسـب مستوياتها التـي تعكس مدى تعقيدها ومدى كفاءتها في الترجمة إلى الأصناف الآتية بشكل تقريبي (٩و٢٦):

المسـتوى الأدنـى: تستدعـي الترجمة الآليـة البدائية ترجمة كلمـة لكلمة دون أيـة معالجة آلية قبل ذلـك، ويمكن بعد ذلك للمترجم البشـري أن يفهم من المادة الناتجـة المقصود مـن العبارات فيقوم بإعـادة صياغتها بشـكل جديد بحيث يعيد ترتيـب الكلمات والعبارات ويولـد الجملة المطلوبة ويقوم بالتأكـد من دقة وسلامة الجملـة. وهذا المستوى يحتـاج إلى معجم ثنائي اللغة ضخـم ومعجم للكلمات التي تحتاج أكثر من كلمة تقابلها.

المستوى التالي الأعلى من المستوى الأدنى: في هذا المستوى تجرى بعض المعالجات الصرفية لغرض التعامل المخفي مع الكلمات. ويمكن اعتبار الترجمة بالأمثلة ضمن هذا المستوى بوجود ذخيرة من جمل وعبارات متوازية.

المسـتوى المتوسـط الأول: يحتاج هذا المستوى إلى شجرة إعراب للجملة (محلل نحوي) باللغة المصدر ومن ثم تسقط على اللغة المترجم إليها.

المسـتوى المتوسـط الثـاني: في هذا المسـتوى هناك حاجة للتمثيـل الدلالي بين اللغات، فمثلًا إذا قلنا "رأسي يؤلمني" بالعربية، فترجمتها بالإنكليزية هي "I have headache" والتي تترجم حرفيا إلى "لدي صداع". وعلى هذا فمن الضروري فهم المعنى وتمثيله بشكل سليم، لكي يمكن إيجاد المرادف له في اللغة الأخرى.

المستوى الأعلى: لا يزال البحث في هذا المستوى يجري، وذلك للأخذ بعين الاعتبار الأساليب البلاغية العميقة في اللغتين، وتكاد تعاني كل برامج الترجمة الآلية حاليًا من قصور في هذا الجانب (٢٢).

وفي كل الأحوال على برنامج الترجمة الآلية أن يقوم بعمليتين رئيسيتين:

١. **النقل:** أي عملية إيجاد ما يقابل كلمات وعبارات وجمل النص الأصلي، وكذلك إيجاد التراكيب النحوية المقابلة للتراكيب الواردة فيه.

٢. أما العملية الثانية فهي صياغة الجمل الناتجة من عملية النقل السابقة صياغة صحيحة صرفيًا ونحويًا وأسلوبيًا، مثل صياغة: (عامل) + (جمع) في صورة (عمال) و (معلم) + (جمع) + (حالة النصب أو الجر) في صورة (معلمين)، وكذلك وضع الصفة في العربية بعد الموصوف ومراعاة قواعد المطابقة اللازمة (١و٢).

وفيما يأتي تفصيل لأهم أساليب الترجمة الآلية المستعملة اليوم. ونشير إلى أن كل هذه الأساليب ما زالت قاصرة إن استعملت لوحدها، لذلك فمن الضروري للارتقاء بالترجمة الآلية اعتماد أكثر من أسلوب بشكل متوازي لرفع مستوى الترجمة خاصة لفك الغموض (٢١و٢٤).

الترجمة باستعمال لغة وسيطة

تستند لغة التشبيك العالمية (٢٧) إلى تحويل اللغة المصدر إلى لغة تمثيل للنص لا يعتمد على أية لغة، ومن ثم يمكن ترجمة هذه الرموز إلى أية لغة أخرى، وهذا يعني أن هذه الطريقة يمكن أن تخدم الترجمة إلى لغات متعددة بالتوازي، بينما تحتاج الطريقة الإحصائية التي سيرد ذكرها فيما بعد إلى جهد وذخيرة لغوية متوازية بين كل لغتين يراد الترجمة بينهما. إن التحدي الذي تواجهه هذه الطريقة هو تكوين لغة تمثل المقصود باللغات كلها دون لبس لكي يمكن الترجمة منها. يحتوي برنامج الترجمة بهذه الطريقة على ثلاثة أجزاء هي:

الوحدات اللغوية والتي تضم المعاجم التي تحتوي على كلمة عالمية تمثل المفردات العالميـة وتضم القواعد النحوية لتحويل اللغـات الطبيعية إلى هذه اللغة العالمية بما في ذلك من علاقات وأوصاف وملصقات وكل ما تحتاجه اللغات الطبيعية لكي تكون عبارات فصيحة وواضحة ومفهومة في اللغة المراد الترجمة إليها، وهذا ما يستدعي وجود قاعدة معرفة هرميــة الشكل للمفاهيم الموجودة في اللغات الطبيعية. إن هذه اللغـة هي لغة اصطناعية يمكن بوساطتها العبـور نحو اللغة الهدف وتتمثل كفاءتها في دقة تمثيل اللغة المصدر ومدى إمكانيتها لتحويل ذلك التمثيل إلى اللغة الهدف.

إن الفكرة وراء الترجمة الآلية المستندة إلى التحويل (٢٨) هي أن يكون هناك تمثيـل وسيط يلتقط معنى الجملـة الأصلية بحيث يكون وعاء للمعرفة لا يعتمد على لغـة معينة ولا علـى حقل معين لغرض توليـد الترجمة الصحيحـة. ويمكن تحسين الترجمـة بصياغـة المعرفـة اللغوية وتزويـد الحاسـوب بقواعد كافيـة للتعامل مع الخصائصـ اللغوية. إن قوة هذه الطريقة الحقيقية هـي في إمكانية تمثيل المعرفة بهذه اللغة الوسيطة.

تعتمـد الطريقة على تمثيـل النص من اللغة المصدر بشـكل شبكة نحوية كلية باستعمـال لغة تسمى لغــة التشبيك الكلية والتي يمكن بعد ذلك تحويلها إلى أية لغة أخرى (٢٧).

وقـد جربت هـذه الطريقة على نحـو ١٠٠٠ صفحة من موسوعـة أنظمة دعم الحيـاة إلى اللغـات الستة التي تدعمها وثائق اليونسكو وكانـت النتائج لا بأس بها. وعلـى هذا فحينما يكون المطلـوب ترجمة نص لعدد من اللغات فـإن التحويل بهذه الطريقة قد يكون خيارًا مفضلًا (٣٠).

الترجمة الآلية المستندة إلى الإحصاء

تحتاج هذه الطريقة الحديثة لجمع أكبر ما يمكن من ذخيرة لغوية (corpus) والقيـام بجهد إحصائي كبـير عليها لكي تصبح مناسبـة للاستخدام في الترجمة

الآلية. وبالطبع فإن الذخيرة المترجمة هي بالأساس مترجمة من قبل مترجمين من البشر، وهي بذلك تستخلص خبرات البشر للإفادة منها في الترجمة الآلية (١).

تعتبر الذخيرة اللغوية لأية لغة مرتكزًا أساسيًا اليوم للباحثين في اللغات الطبيعية. فالخزن على الحواسيب جعل جمع وتصنيف وتحليل الذخيرة اللغوية ميسورًا. وهذه الذخيرة تمكن الباحثين من إطلاق الوصف على خصائص اللغة وعلى النحو فيها ودراستها تاريخيًا وتغيرها مع الزمن (١٣).

تجمع الذخيرة اللغوية المناسبة (يجب أن تكون من الضخامة ما فيه الكفاية) بين محتوياتها غالبية الكلمات الشائعة في اللغة وغالبية التعابير اللغوية والتراكيب النحوية والصرفية. وعلى هذا فيتوقع منها أنها تمثل اللغة بشكل تقريبي. ونظرًا لاعتماد هذه الطريقة على الذخيرة ثنائية اللغة (أو متعددة اللغات)، فيجب أن تكون هذه الذخيرة شاملة وواسعة ودقيقة، وهو أمر في غاية الندرة. وتتوفر مثل هذه الذخيرة اليوم في اللغات العالمية الأخرى غير العربية لضخامة ما يتوفر من كتابات محوسبة كالصحف والمجلات والكتب المطبوعة والمقالات المنشورة على الإنترنيت والإعلانات ونشرات الشركات التجارية وهي تمثل اللغة، ومن ثم تستفيد الترجمة الآلية من مثل هذه الذخيرة (١٥و١٦).

إن الحجم الأدنى للذخيرة التي يمكن الاعتماد عليها لغرض الاستعمال في الترجمة يبلغ مئات الملايين من الكلمات. وبهذا الصدد لا تزال الذخيرة اللغوية المتوفرة باللغة العربية مع ترجماتها للغات الأخرى محدودة.

هناك عدة أنواع من الترجمة الإحصائية، فمنها المباشرة التي تستند إلى معجم عبارات للغتين. أما الترجمة الإحصائية غير المباشرة فيجري فيها القيام بعمليات تحويل لغوية عديدة للإفادة من الذخيرة كالتحويلات الصرفية والنحوية والدلالية وحذف السوابق واللواحق والقيام بعمليات إحصائية واستعمال الأنظمة الخبيرة. كما أن هناك الترجمة الإحصائية بالأمثلة التي تحتاج للقيام بإعراب

الجملـة بلغتها الأصلية ثـم القيام بعدد من التحويلات بإعـادة السلسلة الظاهرية في اللغـة الأصلية لكي توافق التسلسل المطلـوب في اللغة المترجم إليها. وتجرى هذه العمليـات على الجزء من الذخـيرة المستعمل للتدريب وكذلك على النصوص المراد ترجمتهـا ويحتاج ذلك لحل مشكلـة تقابل عدة كلمات في لغـة مقابل كلمة في اللغة الأخرى وبالعكس (١٧).

تستنـد عملية تهيئة الذخـيرة حاسوبيًا على استخـراج الكلمات المكررة وغير المكـررة في الذخـيرة (٣١و٣٢) ثـم بعـد ذلـك الحصول علـى قائمة بـكل كلمتين متعاقبتـين ثم قوائـم العبارات المتقابلـة بين اللغات المـراد الترجمة منهـا وإليها. وبالطبـع فـإن هذه الطريقة بأسلوبهـا المبسط لا تعطي ترجمـة دقيقة وهي بحاجة إلى معلومات نحويـة وصرفية لزيادة الدقة. وقد استعملت هذه الطريقة في ترجمة محاضـر البرلمـان الكندي بين اللغتـين الفرنسية والإنكليزيـة فوجدت ناجحة لحد كبـير، وذلـك بسبب أن موضوع التحاور في أروقة البرلمـان ذو نمط معين وبأساليب متعارف عليها بين المتحاورين (١).

إن الخطـوة التالية في معظم برمجيات الترجمة الآلية المستندة إلى الإحصاء بعـد استكمال الذخيرة اللغويـة المتوازية هي توليد الكلمـة المقابلة بين اللغتين وحل مشكلة عدم تساوي عدد الكلمات بين اللغتين (١٨).

تعتمـد دقـة الترجمـة الإحصائية على نـوع ودقـة الذخـيرة المستعملة وعلى البرمجيات المستعملة فيها (٢١). ويجدر بالذكر أن تغير اللغة والمفردات المستعملة بمضـي الزمـن وضخامة الذخـيرة المطلوبة يستوجـب وجود معجـم تأريخي للغة. كمثـال علـى ذلك فإن كلمـة "السيارة" حين تـرد في نص مكتوب قبـل قرون تعني غـير "السيارة" المكتوبة في نص مستعمل اليوم. وكذلك فإن استعمال بعض أسماء الأعـلام يحتـاج وجـود قرائـن في البرنامج لكـي يقـوم بترجمتها بمعناهـا وعدم اعتبارهـا أسماء أعلام. فمثـلا كلمة: "أحمد" في جملة "أحمـد الله على نعمائه" هـي ليست إسم علم. ومن الصعوبة بمكان علـى الآلة تحديد هل الكلمة ذات معنى أم اسم علم بدون قرائن كافية.

الترجمة بالأمثلة

الترجمة المعتمدة على الأمثلة تستند إلى ترجمة جمل قصيرة باستعمال ذخيرة لغويـة ضخمـة مبوبة على مستـوى العبارات وقـد استفيد من وثائـق الأمم المتحدة متعـددة اللغات بهذه الطريقـة. لا تعتمد هذه الطريقة من الترجمـة على الهيكلية، حيـث أن الأمثلـة تخـزن بشكلهـا السطحي مـع تعامـلات صرفية وتجزئـة للكلام محـدودة، حيـث تقابل كل جملـة مع ترجمتهـا باستعمال عمليـات صرفية محددة. وتجـري عملية المقابلة بالرجوع إلى خوارزمية تحـدد المعاني التقريبية للكلمات بدل استعمـال معاجـم ثنائية اللغة. وأهم ما في هذه الطريقـة أنها تحاكي الطريقة التي يفكـر بها المترجم البشـري، فالذخيرة التي تستند إليها هـذه الطريقة هي القوائم متعددة اللغات للكلمات والعبـارات والجمل المتوازية. وتقوم الترجمة بإيجاد المقابل للنصـ المطلـوب ترجمته من القوائـم المتوفرة. وهذه العملية تسمـى التحويل، تليها عملية دمج العبارات المختلفة لتوليد النص النهائي باللغة الهدف (١).

إن الترجمـة بالأمثلـة لا يمكن أن تعتـبر كافية لإتمام عمليـة الترجمة كاملة، لكنهـا يجب أن تتم مع طرائـق أخرى للترجمة بحيث يكمل بعضهـا بعضًا. وتستند هذه الطريقة إلى إجراء أبحاث على الذخيرة اللغوية لاستخلاص الأمثلة والعبارات الشائعـة المتقابلة بين اللغتين المراد الترجمـة بينهما، إضافة إلى استخلاص قوالب نحويـة بـين اللغتين لتعويضـ الكلمات المقابلة بـين اللغتين. هـذه الطريقة فعالة في العادة للترجمة بـين اللغات التي تعود لأصول لاتينية مثلًا. لكن الأمر يزداد صعوبة بين اللغات التي تعود إلى عوائل مختلفة.

الترجمة الآلية واللغة العربية

تمتـاز اللغـة العربية بميزات خاصـة مختلفة عن اللغات الأخـرى، وفي الوقت نفسـه تجمعهـا مع اللغـات الأخرى ميزات عامـة، لذلك يمكن الاستفـادة من هذه الميـزات العامة مع اللغات الأخرى والبناء عليهـا لتغطية خصوصيات اللغة العربية

سـواء في الترجمة من العربيـة أو الترجمة إليها. فمن الأمور المشتركة بين عدد من اللغـات تعـدد المعاني للكلمة الواحـدة ووجود جمل ليس فيها فعـل. كما تحوي اللغة العربيـة على عوامل صرفية عديدة منها السوابق واللواحق والأواسط وتغيير صيغة الكلمة. هذا يعني أن أية ترجمة آلية يجب أن تحوي تحليلا قويًا للوصول إلى الجذر.

إن الترجمـة الجيدة يجب أن تقتنص المعنـى بكفاءة، ونعني بالكفاءة الوصول إلى المعنـى المقصـود بدقة باستخـدام أكبر ما يمكن من قرائـن. ،وهذا ما تفتقده أنظمة الترجمة الآلية بالوقت الحاضر، وهو ما يضع عبئًا على من يقوم بالتصحيح بعـد الترجمة، والذي قـد يكون أكثر مما يمكن أن يبذل بالترجمـة البشرية أحيانًا إذا كانت الترجمة سيئة.

أمـا خصوصيـات اللغة العربيـة ذات العلاقـة بالترجمة اليـوم فمنها غياب التشكيل رغم أنه جزء رئيسي من النص المنطوق، مما يفرض على القارئ افتراض تشكيـل من عنـده بما يملكه من ثـروة لغوية. وهو ما يقوم بـه المترجم كذلك، مما يستدعـي إيجاد بديـل لذلك في الترجمة الآليـة (١٠). ويزيد الأمـور تعقيدًا ندرة استعمال علامات الوقف والفواصل في النصوص العربية.

كمـا تمتاز اللغة العربية بدمج السوابـق واللواحق والإدغام والإقلاب وغيرها. أمـا الجمل الطويلة في اللغة العربية فتحتـاج إلى عمليات حاسوبية معقدة لتحليلها مما يؤدي إلى أن تصبح المعالجة بطيئة. ومن العمليات التي تساعد في ذلك تقسيم الجملة إلى عبارات المضاف والمضاف إليه والجار والمجرور وأشباه الجمل (١١).

أما من الناحية التأريخية فإن التغييرات التي صاحبت اللغة العربية كانت أقل بكثـير من لغات أخرى كالإنكليزيـة التي شهدت تطورًا هائلًا خلال القرون الأخيرة بالمقارنة مع اللغة العربية التي لا يزال تراثها المكتوب منذ صدر الإسلام ولحد الآن لم يحدث عليه تغيير جوهري في المفاهيم العامة.

ولغرض أن تكون الترجمة الآلية مقبولة يجب القيام بعدة عمليات على النص العربي برمجيًا، منها على سبيل المثال لا الحصر ما يأتي (١٢):

- إضافة الرموز للنص كالفوارز والفواصل والنقاط ومعالجة الهمزات والتفريق بين الهاء والتاء المربوطة والتشكيل للأحرف التي فيها لبس.
- مراعاة وجود السوابق واللواحق وتحديد أجزاء كل منها إن وجدت كألف لام التعريف وحروف العطف وحروف الجر المتصلة والضمائر وعلامات الإعراب وتحديد جذر جذع الكلمة ووزنها الصرفي إذا احتيج لذلك.
- مراعاة قواعد تغيير كتابة الكلمة بعد إلحاق بعض الضمائر المتصلة مثل كلمة "مكتبة – مكتبتهم" و"أعلى – أعلاه".
- تحديد فيما إذا كان العدد المتضمن ضمن الكلمة مفردًا أو مثنى أو جمعًا وذلك من خلال علامات الجمع الواو والنون والياء والنون وعلامات المثنى من خلال الألف والنون والياء والنون وغيرها.
- تحديد المعنى عندما يكون هناك وجود لأكثر من احتمال في تحديد الكلمة المراد ترجمة معناها مثل الكلمات التي تعطي أكثر من معنى باختلاف التشكيل كأن يكون الفعل مبنيًا للمجهول أو للمعلوم أو أن يكون اسمًا أو فعلًا مثل: ذَهَبَ و ذَهَبٌ (١).
- التحليل النحوي لتحديد التركيب الدقيق للعبارة أو الجملة حيث يقوم البرنامج بتطبيق قواعد منطقية بهدف الوصول إلى الترجمة باللغة الهدف.
- الربط بين تركيب الجملة ومركباتها الدلالية لغرض الوصول إلى الترجمة الهدف.
- التركيب المنطقي بتمثيل الجملة الأصلية باللغة الهدف.

- تحسين الجملة باللغة الهدف نحويًا وصرفيًا ودلاليًا.

الأبحاث في الترجمة من اللغة العربية وإليها

اهتمت في الآونة الأخيرة كثير من الجامعات الأمريكية والأوربية والعربية بمعالجة اللغة العربية حاسوبيًا والترجمة منها وإليها. وتتوزع الأبحاث على النحو والصرف والدلالة والترجمة وجمع الذخيرة اللغوية وأساليب تكوين المعجم المحوسب وعملية الاستفسار باللغة العربية وتكوين الخلاصات وغيرها (١).

إن الكلمات العربية المركبة غالبًا ما تتكون من سوابق ولواحق، وحيث إن عملية فصل هذه اللواصق فيها كثير من الإبهام، لذلك هناك حاجة للرجوع إلى المعنى بالإضافة إلى القواعد الصرفية والنحوية. كما أن عملية فصل اللواصق تزداد صعوبة في العربية بسبب غياب التشكيل عادة.

تقع عملية فصل اللواصق في ثلاثة أصناف: صنف يستند إلى قواعد وصنف يستند إلى الإحصاء وصنف يعتمد المعجمية كأساس له. الصنف الذي يعتمد القواعد يستند على المعرفة اللغوية المأخوذة من الخبرة البشرية بحيث تحول هذه المعرفة إلى قواعد وهذه القواعد معروف أن الوصول إليها ليس بالسهولة بمكان، كما إن الكثير من هذه القواعد ليست حدية ولها استثناءات. أما الصنف الذي يعتمد على الإحصاء فيحتاج إلى تدريب وبعض المعرفة البشرية لبنائه.

لقد وجد أن القيام بعملية رصينة في فصل اللواصق ذو تأثير أفضل بكثير من القيام بالعملية بشكل سطحي. ويمكن أن يتحقق تحسن آخر باستعمال أكثر من وسيلة من وسائل فصل اللواصق.

إن العوامل التي تحكم ترجمة التعابير المركبة متعارف عليها لدى اللسانيين، وحيث إن معظم التعابير المركبة تكون مركبة من أسماء لذلك فهي تخضع لتصريف الأسماء في اللغة، أما استعمال حروف الجر في اللغة العربية وتلازم بعضها مع

الأسمـاء فإنـه يجعل التعامل معها غـير ممكن بشكل مباشـر (١٦و١٧). ويلاحظ أن اللغـات المختلفة يمكـن أن تحتوي كلمـات متشابهة التسلسل وكلمـات متعاكسة التسلسل وكلمات متشابهة التهجئة.

ومما يجب معالجته مشكلـة المتلازمات اللفظية التي يمكن تقسيمها إلى أربع فئات حسب درجة التّلازم، فالمتلازمات اللّفظية المفتوحة الحرة مثل "بادئ الأمر" يمكـن استبدالها بـ "أول الأمـر"، فاستبدال كلمة "بادئ" بمرادفها "أول" لم يخل ذلـك بالمعنى. والمتلازمات اللّفظية المقيـدة وهي أن ترد لفظتان معًا بحيث تستعمل إحداهمـا بمعناهـا الحرفي وتستعمل الأخـرى بمعناها المتخصص الذي يمكن أن يكـون مجازيـا مثلًا "يشقّ طريقه". ففـي هذا المثال، لا يعني الفعـل "يشقّ" معناه الحرفي أي "أن يقسم المرء شيئًا إلى شطرين" بل معنىً مجازيًا هو "أن يسلك المرء طريقًـا صعبًـا". وهناك العبارات المسكوكة المجازية مثـل قولنا "والليل إذا تنفس" والعبـارات المسكوكة المحضة مثل "وضعت الحرب أوزارها"، بمعنى انتهت وتوقفت إذ لا يمكـن في هـذه العبارات إبدال كلمة بكلمة أخرى أو تقـديم إحداهما أو تأخير أخرى (٧).

ومما يجب تكوينه المعجم العربي المحوسب الذي ينبغي أن يحصر جميع المعاني للكلمـة ودراسـة مدى شيـوع الكلمات ودراسة تكـرار الكلمة كمـادة وكفئة معجمية ودراسـة استخـدام المرادفات وشكل الكلمـة وفق حالتها الإعرابيـة ودراسة الكلمة وفـق اشتقاقها الصرفي ومعرفة مدى السلامة والصحة اللغوية ومعرفة المتلازمات اللغوية (٧). كما أن المعجم المحوسب يجب أن يحوي رموزًا خاصة لتصريف الكلمة ومعلومـات أخرى عنها تدرج فيهـا مفردات اللغة بالتفصيل بحيـث يكون بالإمكان الإفادة منها حاسوبيًا.

محاولات الترجمة الآلية من اللغة العربية وإليها

قامـت محاولات عديدة للبدء بترجمة آلية مـن اللغة العربية وإليها. وقد أثمر

بعضهـا بتكوين أنظمة ترجمة آلية، بينمـا أصبِح البعض الآخر طي النسيان. وفيما يأتي بعض من هذه المحاولات (١):

- نظـام "المترجم العربي" الـذي طورته شركـة ATA في لندن، وبرنامجا مصغرًا أسمته "الوافي" (٦).
- نظـام "الناقل العربي" الذي طورته شركـة سيموس في باريس للترجمة بين العربية وكل من الإنكليزية والفرنسية (٣٥).
- نظام شركة أبتك Apptek (٣٦).
- نظام سيستران Systran ويترجم النظام هذا الآن بين١٥ لغة. لقد ابتدأ تطويـر سيستران لمترجـم من العربيـة إلى الإنكليزية في عـام٢٠٠٢ وله موقـع للترجمة الآن (٣٧)، وقد امتاز سيسـتران نتيجة استعماله المكثف للمعاجم بقابليته لترميز مكونات اللغة التفصيلية بميزات نحوية وصرفية دقيقـة. لكن دراسة لفحص المتلازمـات اللفظية (التي تشكل عائقًا أمام المترجم والمتعلم) على نظام سيستران وجدت أن النظام أخفق في ترجمة معظم المتلازمات اللفظية (٧).
- شركـة ألبسـ Alps: لديها برامج للترجمة بين عـدد من اللغات، وتطبق مبدأ الترجمة التحاورية (٣٨).
- موقـع المسبار وهو موقع يهتم أيضا بالترجمـة الآلية من الإنجليزية إلى العربيـة وبالعكسـ، ويمتاز بالسهولـة والمرونة عند استخدامـه إذا كانت العبارات بسيطة (٣٩). ويرتبط موقعه مع المترجم العربي والوافي.
- موقع freetranslation ويترجم بين٤٣ لغة (٤٠) بينها اللغة العربية.
- لقد وظفت شركة صخر لبرامج الحاسب محركًا للترجمة الآلية الخاص بها (٨) في دعم موقع الترجمة (٤١)، والموقع متوقف الآن.

- برنامج شركة جوجل: وهو برنامج مجاني يستند إلى الترجمة الإحصائية من ذخيرة لغوية مأخوذة من الإنترنيت ويتعلم من أخطائه، فإذا ما ترجم جملة خاطئة وأخبره المستخدم أن الترجمة خاطئة وأن المفروض أن تكون بشكل آخر، فإنه يخزن هذه المعلومات ويستعملها في المستقبل بشكل أصح. عدد اللغات التي يترجم منها وإليها تبلغ ٨٠ لغة (٤٢) بينها اللغة العربية.

- ترجمة بينغ: من شركة مايكروسوفت. وهي تُستخدم من قبل فيسبوك وياهو لترجمة التعليقات. تستخدم بينغ "الترجمة الآلية الإحصائية"، غير أنها تختلف عن ترجمة جوجل في طريقة تنفيذها أو البيانات التي تعتمد عليها. وتشمل ترجمة بينغ ٤٥ لغة بينها اللغة العربية، وتسمح خاصية الترجمة التعاونية للمستخدمين بأن يصححوا أو يحسنوا الترجمات (٤٣).

مقارنة بين بعض برامج الترجمة الآلية من اللغة العربية وإليها

سنقارن بين ثلاث برمجيات للترجمة الآلية هي جوجل وسيستران وبينغ المذكورة مواقعها أعلاه. وقد اخترنا جملة عربية لترجمتها ومقارنة ترجماتها إلى الإنكليزية. ثم ترجمنا الجملة إلى الإنكليزية وأدخلت تلك الترجمة للأنظمة الثلاثة للحصول على الترجمة العربية. وقد أعيدت التجربة مرتين في شهري أيلول (أشير لها بالرمز (×)) وكانون الأول (أشير لها بالرمز (××)) من عام ٢٠١٤ ويلاحظ حدوث اختلاف في كل الترجمات من الإنكليزية إلى العربية في المواقع الثلاثة، بينما الترجمة من العربية إلى الإنكليزية اختلفت في ترجمة جوجل وبقيت نفسها في سيستران وبينغ. ويلاحظ عدم الدقة والركاكة في الترجمات الآلية كلها تقريبًا. كما يلاحظ أن الترجمة قد تختلف بمضي الزمن.

النص الأصلي بالعربية	حصل تقدم هائل للترجمة الآلية خلال السنوات الماضية خاصة بين اللغات الأوربية ذات الأصول اللاتينية لما بينها من مشتركات
ترجمة جوجل إلى الإنكليزية	Got tremendous progress of the translation mechanism during the past years, especially among European languages with Latinos because of the participants, including (*) Progress has been made tremendous machine translation during the past years, especially among European languages with Latin asset to them from participants (**)
ترجمة سيستران إلى الإنكليزية	Huge progress for the automatic translation during the last years collected especially between the European languages self of the Latin origins for what between her from joint (*) & (**)
ترجمة بينغ إلى الإنكليزية	Advances of machine translation over the past years, especially between the European languages with Latin origins to their subscribers (*) & (**)
ترجمة النص بالإنكليزية	A tremendous progress occurred during recent years between European languages due to their common Latin origin.
ترجمة جوجل إلى العربية	حدث تقدم هائل خلال السنوات الأخيرة بين اللغات الأوروبية بسبب الأصل اللاتيني المشترك (×) حدث هناك تقدما هائلا خلال السنوات الأخيرة بين اللغات الأوروبية بسبب الأصل اللاتيني المشترك (××)
ترجمة سيستران إلى العربية	تقدم هائل وقع أثناء سنوات الأخيرة بين لغات أوربيّ واجب إلى أصلهم لاتيني عادي (×) تقدم هائل وقع أثناء سنوات الأخيرة بين لغة أوربيّ واجب إلى أصلهم عاديّ لاتينيّ (××)

ترجمة بينغ إلى العربية	التقدم هائل الذي حدث خلال السنوات الأخيرة بين اللغات الأورو بية بسبب أصلها اللاتيني الشائعة (×) التقدم هائل الذي حدث خلال السنوات الأخيرة بين اللغات الأورو بية بسبب أصلهم اللاتينية الشائعة (××)

في حال الترجمة من العربية إلى الإنكليزية، ينصح (٤٤) باستخدام العربية الفصحى البسيطة الشبيهة بما قد يوجد في المواقع الإخبارية، والابتعاد عن التعبيرات البلاغية. كما ينصح باستخدام بنية الجملة التي تبدأ بالاسم ثم الفعل كما في بنية الجملة الإنكليزية، مثلا "الرجل جاء" بدلًا من "جاء الرجل". فحين تكون الجمل العربية طويلة قد يَتْبَع الفعل أكثر من عشر كلمات قبل ورود الفاعل، مما يجعل من الصعب على الترجمة الآلية أن تنقل الفعل إلى مكانه الصحيح في الجملة الإنكليزية، بل إن كثيرًا من أنظمة الترجمة الآلية تحذف الفعل بالكامل.

برامج مساعدة للترجمة

تخدم المعالجة الآلية للغة العربية بالإضافة إلى الترجمة الآلية: التلخيص الآلي والتوليد الآلي للغة واستخلاص المعلومات واسترجاع المعلومات والإجابة على الأسئلة والتنقيب في النصوص وتحويل الكلام المنطوق إلى مكتوب وتحويل النص إلى كلام منطوق والتعرف الضوئي على الحروف ... وهناك أبحاث جارية لكل هذه المواضيع. ومن البرامج المتوفرة الآن: برنامج التشكيل والتصحيح الآلي (٤٥) Arab Diac ومجموعة برامج صخر للتدقيق الإملائي Sakhr Corrector وبرنامج التشكيل الآلي (٤٦) Automatic Diacritizer وبرنامج صخر للتحليل الصرفي (٤٧) وبرنامج استخراج الكلمات المفتاحية (٤٨) Sakhr Keywords Extractor ومحرك صخر للتصنيف الشجري "سراج" ومحرك صخر للتلخيص الآلي وبرامج تنقيب ووصف الكلام (٤٩) Part of Speech Tagging . وهناك برامج التحليل الصرفي لتحديد جذور الكلمات والمعلومات الصرفية والنحوية ومن هذه البرامج:

برنامـج (٥٠) ArabMorpho حيـث يستخـدم ذخيرة للألفـاظ وتحليل إحصائي لفـك الغموض ومحلل (٥١) Beesley Xerox وهـو مولد ومحلل صرفي مساعد في التعليم وبرنامج باك وولـتر للتحليل الصرفي (٥٢) وبرنامج التنقيب في النصوص للوصول إلى الأفكار الأساسية وبرنامج محرك البحث النصي العربي (٥٣) Swift وبرنامج Arab Dictions الذي يقوم بتحليل الكلمات العربية إلى وحداتها الصرفية ثم يربط كل الأشكال الصرفية بمادة المعجم المقابل لها وفق الجذر (٥٤)، وبرنامج معالجة الكلام المنطوق Arab Talk الذي يحول النصوص المكتوبة إلى كلام منطوق معتمدًا على تشكيل آلي (٥٥)، وبرنامج إبصار للمعاقين بصريًا والمكفوفين (٣٣).

ترجمة الكلام المنطوق

تتوفر في الأسواق بعض القواميس الصوتية لترجمة عبارات بين لغات متعددة بينهـا اللغة العربية. وكثيرًا ما تحتوي عبارات باللهجـات العامية. ويمكن استعمال بعض الأنظمـة الصوتية في حجز الفنادق والحجز على الخطوط الجوية والمشاركة في المؤتمـرات وطلـب الطعام مـن المطاعـم والاستفسار عن اتجـاه السير والحجز لـدى العيادات الطبية والمستشفيات واستئجار السيارات وغيرها (١٩). إن مشاكل ترجمـة الـكلام المنطوق تزيد علـى مشاكل النصـ المكتوب، وذلـك لأن الكثير من الـكلام المتـداول يحوي أخطاء نحوية أو جملًا ناقصـة. ولغرض تكوين ترجمة آلية فوريـة، يجب وضع قواعد عامة ذات مرونة لا تلتـزم بقواعد اللغة بصرامة، وذلك للأخـذ بعين الاعتبار تحديـد نطاق الكلمات التي يستعملهـا المتكلم وطريقة نطقها وسرعـة النطق وطول الجملة وصيغ التوقف بين الجمـل. وهناك محاولة لاستعمال الرسائـل القصيرة علـى الهاتف النقـال للترجمة وتمتاز بأنهـا قصيرة ومحدودة المعجم، وبأنهـا قابلة للتطور والتوسع (٢٠). ولا تزال الأبحاث لمثل هذه الأنظمة في مراحل التطوير.

الاستنتاجات والمقترحات

الترجمة الآلية مـن اللغة العربية وإليها إحدى الوسائل التي تكمن فيها خدمة كبـيرة في نهضة الأمة وأجيالها المستقبليـة وذلك لأن اللغة العربية اليوم ليست هي اللغـة العالمية للعلوم والتقنية، فما يصـدر في اللغات الأخرى وخاصة الإنكليزية من أبحاث ودراسات وكتب ومؤلفات أخرى يحتاج للترجمة للغة العربية. كما أن اللغات التـي يتكلم بهـا المسلمون بحاجة إلى التراث العربـي الإسلامي غير المتوفر في تلك اللغات وهناك حاجة ماسة لترجمته. لذلك فالترجمة الآلية من اللغة العربية وإليها ليست من الكماليات بل هي من الضرورات القصوى لنهضة الأمة وعودتها لمكانتها بـين الأمم. وأول ما تحتاجه الترجمة الآلية هو وجود معجم عربي محوسب وتكوين ذخيرة متعددة اللغات مع العربية لتساعد في الترجمة الآلية المستندة إلى الإحصاء ودعـم الأبحاث اللغوية المتعلقـة بالترجمة الآلية من اللغة العربيـة وإليها. ولغرض الوصول إلى ذلك هناك حاجة إلى القيام بحملة توعية للقيادات السياسية والعلمية والجهات الداعمة للبحث العلمي على أهمية البحث العلمي في حوسبة اللغة العربية مـن قبل المجامـع اللغوية العربية وأقسـام الحاسوب واللغة العربيـة واللسانيات في الجامعـات العربية كما ينبغي توجيه الأبحاث نحـو التطبيق العملي وتكوين قيادات بحثية في أقسام اللغة العربية ذات خلفية حاسوبية جيدة وفي أقسام الحاسوب ذات خلفيات جيدة باللغة العربية لكي يكون التواصل والبحث العلمي على أتم وجه (١).

المصادر

المصادر العربية

١. محمد زكي خضر، اللغة العربية والترجمة الآلية – المشاكل والحلول - المؤتمر الحادي عشر للتعريب – عمان / الأردن ١٢–١٦ تشرين الأول ٢٠٠٨م.

٢. محمود اسماعيل صالح (الصيني) - الحاسوب في خدمة الترجمة والتعريب: http://www.wata.cc/forums/showthread.php?5421- الحاسوب–في–خدمة –الترجمة–والتعريب–

٣. محمد بن محمد الغزالي أبو حامد - المستصفى في علم الأصول- ١٨

٤. محمد الصرايرة - اللغة العربية والترجمة الآلية – محاضرة في مجمع اللغة العربية الأردني - الموسم الثقافي التاسع عشر ٢٠٠١م

٥. أبو الحجاج محمد بشير – المعالجة الآلية للغة العربية جهود الحاضر وتحديات المستقبل - مجلة لغة العصر المصرية – ٢٠٠٩

٦. عيدان، عدنان، طارق إبراهيم، الترجمة الآلية من اللغة الإنكليزية إلى اللغة العربية: تجربة شركة أي–تي–أي . لتقنية البرامج، مركز دراسات الوحدة العربية، ص ٢٨٩

٧. آمنة فاطمة الزهراء – إشكالية حدود الترجمة الآلية: ترجمة نظام "سيستران" للمتلازمات اللفظية – رسالة الماجستير كلية الآداب و اللغات - قسم الترجمة - جامعة منتوري –قسنطينة–الجزائر

٨. عبـد الحميد بن العزلان، برنامج "عجيـب" في الترجمة الآلية من اللغة الإنجليزيـة إلى اللغة العربية – رسالـة ماجستير – الجامعة الإسلامية العالمية – ماليزيا – أغسطس ٢٠٠٥

المصادر الأجنبية

1. Thepchai Supnithi، Virach Sornlertlamvanich، Thatsanee Charoenporn، A Cross System Machine Translation COLING-02 on Machine Translation in Asia – Vol 16، Sep. 2002

2. Moustafa Elshafei، Husni Al-Muhtaseb، and Mansour Alghamdi، Machine Generation of Arabic Diacritical Marks، The 2006 International Conference on Machine Learning; Models، Technologies & Applications (MLMTA'06).

3. Habash، Nizar، Bonnie Dorr and Christof Monz.، Challenges in Building an Arabic-English GHMT System with SMT components، Proceedings of the Association for Machine Translation in the Americas (AMTA-2006)، Boston، MA، 2006.

4. Fatiha Sadat، Nizar Habash، Combination of Arabic Preprocessing Schemes for Statistical Machine Translation، Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL، pages 1–8، Sydney، July 2006.

5. Ahmed Abdelali، Jim Cowie and Hamdy S. Soliman، Building A Modern Standard Arabic Corpus، Workshop on computational Modeling of Lexical acquisition. The split meeting Croatia 25th – 28th July 2005.

6. (14) Ali Farghaly، Arabic Machine Translation: A Developmental Perspective، International Journal on Information and Communication Technologies، Vol. 3، No. 3، June 2010

7. (15) Chris Callison-Burch Philipp Koehn Miles Osborne، Improved Statistical Machine Translation Using Paraphrases، Proceedings NAACL-2006

8. (16) Preslav Nakov Hwee Tou Ng، Improving Statistical Machine Translation for a Resource-Poor Language Using Related Resource-Rich Languages، Journal of Artificial Intelligence Research 44 (2012) 179-222

9. (17) Zouhair MAALEJ، Guidelines for the translation of English nominal compounds into Arabic: A computational discourse model، The Tunisian Review of Modern Languages 01/2003; 11:139-167.

10. (18) Saab Mansour and Hermann Ney، Arabic-Segmentation Combination Strategies for Statistical Machine Translation LREC 2012

11. Ying Zhang، Survey of Current Speech Translation Research، Language Technologies Institute، Carnegie Mellon University. citeseer.ist.psu.edu/752690.htm

12. Avinash J. Agrawal، Manoj B. Chandak Mobile Interface for Domain Specific Machine Translation Using Short Messaging Service ieeexplore.ieee.org/iel5/4151644/4151645/04151821.pdf? tp=&isnumber=&arnumber=4151821

13. Nicola Ueffing، Hermann Ney، Word-Level Confidence Estimation for Machine Translation، RWTH Aachen University، portal.acm.org/ citation.cfm?id=1220671

14. Gerd Willée، Bernhard Schröder، Hans-Christian Schmitz، John Hutchins، Computer linguistik: was geht، was kommt? Computational linguistics: achievements and perspectives. Festschrift für Winfried Lenders (Sankt Augustin: Gardez! Verlag، 2002)، p. 159-162] Machine translation today and tomorrow: www.hutchinsweb.me.uk/Lenders- HYPERLINK "http://www.hutchinsweb.me.uk/Lenders-2002.pdf"2002

15. Zhuang Xinglai، The Emerging Role of Translation Experts، in the Coming MT Era، Translation Journal، Volume 6، No. 4 Oct. 2002

16. Fuji Ren and Hongchi Shi، Parallel Machine Translation: Principles and Practice. doi.ieeecomputersociety.org/10.1109/ ICECCS .2001.930184

17. Yasser Salem، Generic framework for Arabic to English machine translation of simplex sentences using the Role and Reference Grammar linguistic model، M.Sc. Thesis، Institute of Technology Blanchards town، Dublin، Ireland، April، 2009

18. Mostafa Aref، Muhammed Al-Mulhem & Husni Al-Muhtaseb، English to Arabic Machine Translation: A Critical Review and Suggestions for Development. The Fourth Saudi Engineering Conference، 1995 Vol. 3، pp421-4

19. Sameh Alansary، Magdy Nagi، Noha Adly، The Arabic Universal Networking Language System، 7th International Computing Conference in Arabic (ICCA)، Riyadh، Saudi Arabia، May 31 - June 2 2011

20. Omar Shirko، Nazlia Omar، Haslina Arshad and Mohammed Albared، Machine Translation of Noun Phrases from Arabic to English Using Transfer-Based Approach، Journal of Computer Science 6 (3): 350-356، 2010

21. W.John Hutchins، Machine Translation، A Brief History، Concise history of the language sciences: from the Sumerians to the cognitivists. Edited by: E.F.K.Koerner and R.E.Asher. Oxford: Pergamon Press، 1995. Pages 431-445

22. Yasser Salem، Arnold Hensman، Brian Nolan، Implementing Arabic-to-English machine translation using the Role and Reference Grammar linguistic model، Proceedings of the Eighth International Conference on Information Technology and Telecommunication (ITT 2008)، Galway، Ireland pp 103-110

23. Abu Shquier، M. and T. Sembok، Word Agreement and ordering in English Arabic machine translation، Proceeding of the International Symposium on Information Technology، Aug. 2008. IEEE Xplore، pp: 1-10

24. Thuy Nguyen & Stephan Vogel، Context-based Arabic Morphological Analysis for Machine Translation، Proceedings of the Twelfth Conference on Computational Natural Language Learning، 2008/8/16 pp 135-142

25. http://www.ebsarfoundation.org/?page__id=151

26. Sabine Hunsicker، Chen Yu & Christian Federmann، Machine Learning for Hybrid Machine Translation، Proceedings of the 7th Workshop on Statistical Machine Translation، pp. 312–316، Montreal، Canada، June 7-8، 2012.

27. http://taln-arabe.blogspot.com/2010/03/mlts.html

28. http://www.apptek.com/products/omnifluent-translate/index.html

29. www.systranet.com

30. Patrick Corness، The ALPS computer-assisted translation system in an academic environment، Proceedings of a conference on Translating and the Computer 14-15 November، London: Aslib، 1986

31. http://www.almisbar.com/

32. www.freetranslation.com

33. http://www.tarjim.com HYPERLINK "http://www.tarjim.com/"/

34. http://translate.google.com

35. http://www.bing.com/translator/

36. https://www.facebook.com/journalismwork/posts/613446758684466

37. http://www.rdi-eg.com/rdi/technologies/arabic__nlp__pg.htm

38. http://www.rdi-eg.com/rdi/technologies/arabic__nlp__pg.htm

39. http://sakhr.software.informer.com/(48) Samhaa El-Beltagy & Ahmed Rafea، KP-Miner: A keyphrase extraction system for English and Arabic documents، Journal of Information System، Vol 34 No. 1، March 2009 pp 132-144(49) http://cl.indiana.edu/~skuebler/papers/arab.pdf

40. http://www.aclweb.org/anthology/W10-3808.pdf

41. www.helsinki.fi/esslli/evening/.../beesley-helsinki.pp

42. http://www.bibalex.org/ica/ar/about.aspx

43. https://www.academia.edu/2547637/

44. http://www.rdi-eg.com/technologies/ArabDiaction.aspx

45. http://arab-talk.com/

تقنيات التعرف الآلي على الكلام المنطوق وتطبيقاتها في القرآن الكريم: واقع وطموح

د.يحيى محمد الحاج (*)

yelhadj@ariscom.org

(*) أستاذ مساعد في جامعة الإمام محمد بن سعود الإسلامية
حصل على درجة البكالوريوس في علوم الحاسب من جامعة القاضي عياض بالمغرب سنة ١٩٩٦ ثم نال درجة الماجستير والدكتوراه في المعالجة المتوازية والموزعة ضمن برامج أوربية مشتركة في سنتي ١٩٩٨ و٢٠٠١. التحق بهيئة التدريس بقسم علوم الحاسب في كلية علوم الحاسب والمعلومات بجامعة الإمام بالمملكة العربية السعودية سنة ٢٠٠٢، وتم تكليفه مؤخرا بالعمل في مركز الحوسبة في المجالات الشرعية والعربية الذي أنشئ حديثا بالجامعة والمجموعة البحثية حول الحوسبة في المجالات العربية والشرعية ARISCOM.

ملخص

تمثل هـذه الورقة بحثـا استطلاعيا حـول تقنيات التعـرف الآلي على الكلام المنطـوق وواقـع استخداماتهـا في مجـال القـرآن الكريم. يتـم في البدايـة تقديم عرضـ مفصل حول تقنيـات التعرف الآلي علـى الكلام، مجـالات تطبيقها بشكل عـام ومفاهيمهـا الأساسيـة وأدواتها التقنية وطـرق بناء تطبيقاتهـا. ومن ثم يتم استعراضـ أهـم الأعمال والجهود المبذولـة في سبيل تطويع هـذه التقنيات لخدمة القـرآن الكريم كرصـد للواقع الفعلي لمـدى الإستفادة منها في مجـال تعلم وتعليم القـرآن الكـريم. وتختم الورقة بلفت أنظـار المهتمين والمعنيـين إلى ضرورة تكثيف الجهـود وتضافرها للإستفادة مـن هذه التقنيات واستغلالها بشـكل أمثل لتسهيل تعلـم وتعليم القرآن الكريم؛ وتشير الورقة إلى نماذجين يمكن أن يكونا نواه لأنظمة فعلية وبيئات تفاعلية للتعلم الذاتي للقرآن الكريم.

مقدمة

يعتـبر مجال التعـرف الآلي علـى الأصوات المنطوقـة أو ما يعـرف أختصارا بالتعـرف الآلي على الكلام من أهم وأقدم المجالات الحاسوبية التي حظيت بجهود بحثيـة معتبرة لكون تواصـل الإنسان مع الآلـة بلغته الطبيعية يمثـل حلمًا طالما قد راوده [١]. ونظـرًا لأهمية الكلام لدى الإنسان والأريحية في التعاطي معه، تزايدت بشكل ملحـوظ الاستخدامات المتعددة لتقنيات التعـرف الآلي على الكلام وظهرت تطبيقاتـه في مجـالات شتى، منها التحكـم عن بعد باستخـدام الهاتف، ومساعدة المعوقـين وذوي الاحتياجات الخاصة، والتعـرف على المتحدث، والتعرف على اللغة، وإدخـال النصوص، والأرشفة ثم البحـث والاسترجاع [١]. وقد برزت في السنوات الأخـيرة توجهـات جديدة تسعى إلى تطبيـق هذه التقنيـات في الأغراض التعليمية وبالـذات في تعليـم القراءة وتصحيـح النطق. وبالرغم من حداثـة هذه التوجهات البحثية، إلا إنها حظيت باهتمام كبير ومتسارع في لغات عدة؛ حيث أطلقت مشاريع عملاقـة في بلـدان مختلفة، نذكر من بينها: مشروعـي "ليسن" [٢-٣] و"كوليت" [٤-٥] لتطوير برمجيـات للتعرف على قراءة الأطفال وتحسين مستواها، ومشروع "سبيسـ" [٦-٧] لتطويـر تقنيات للتعرف على الكلام المنطـوق بشكل يتناسب مع التطبيقـات التعليمية والطبيـة، ومشروع "القارئ الرفيـق" [٨] لمساعدة الأطفال والكبار على تعلم القـراءة، ومشروع "القارئ المساعد" [٩] لمساعدة الطلاب على تحسـين قدراتهم علـى القراءة الصحيحة وزيـادة رصيدهم اللغـوي ورفع مستوى فهمهم.

أمـا علـى مستـوى اللغة العربيـة فالجهود لا تـزال محـدودة، ولم تنتشر بعد الاستخدامـات الفعليـة لتقنيات التعـرف الآلي على الكلام في تعليـم اللغة العربية بشـكل عـام سواء للناطقين بها أو لغـير الناطقين بها على غرار مـا يوجد في لغات أخـرى. ولعـل مـن أهم المجـالات التطبيقيـة لتقنيات التعـرف الآلي علـى الكلام

العربـي المنطوق، تعليم النظام الصوتي للعربية الفصحى لمساعدة المسلمين (عربًا وعجمـا) على تلاوة القرآن بشكل صحيـح سواء لحفظه أو لأداء المناسك والشعائر الإسلاميـة. فالقرآن هو معجـزة الإسلام الخالدة، أنزلـه الله ليتلى ويتدبر ويعمل بـه؛ ونبه الرسول صلى الله عليه وسلم على فضلـه وفضل تعلمه وتعليمه، حيث قال "خيركـم من تعلـم القرآن وعلمه". وقـد ظهر الحرص على حفـظ القرآن الكريم وضبـط تلاوتـه منذ نزولـه على الرسـول صلى الله عليـه وسلـم [١٠]، حيث تولا الصحابة رضوان الله عليهم مهمة حفظه وتحفيظه لجيل التابعين رضي الله عنهم، و انتقل متواترًا عبر الأجيال يتعلمه كل جيل مشافهة عن الجيل السابق. وقد ظهرت الدراسـات المتعلقة بوصف أصوات العربية مـن حيث مخارجها وطرق إخراجها في القـرن الثاني الهجري على يـد الخليل بن أحمد الفراهيدي وتلميذه سيبويه [١١]؛ وتبـع ذلك في القرن الثالث وما بعده تقعيـد علم التجويد وتدوينه كعلم يهتم بتلاوة القرآن الكريم التلاوة الصحيحة [١٢]. وقد أبدع سيبويه ومن تبعه من علماء اللغة في وضـع توصيفـات دقيقة لأصوات العربية، وقد سبقـوا غيرهم من الأمم في ذلك الوقـت. ومع التطور التقني الذي شهده العـالم ظهرت أجهزة ومعدات تقنية مكنت مـن معرفة الجهاز الصوتي لدى الإنسان وما يصـدر عنه من موجات صوتية أثناء الـكلام واستفـادت منها لغـات أخرى؛ ولم يتوقـف الأمر عند هـذا الحد بل برزت تقنيـات حاسوبية تسعى إلى تحديد الصوت المنطـوق والتأكد من سلامة نطقه؛ إلا إن هذه الوسائل التقنية لم تستغل بما يكفي لخدمة اللغة العربية بشكل عام والقرآن الكريم بشكل خاص.

نهـدف في هـذا البحـث إلى استعراضـ تقنيـات التعـرف الآلي علـى الكلام لتقريبهـا من القـارئ العربـي، واستطلاع تطبيقاتهـا الحالية في القـرآن الكريم؛ ونوجـه الراغب في الاستـزادة من المعلومـات المتعلقة بالأجهـزة الصوتية الحديثة وكيفيـة الإفادة منها في ضبط مخارج الحروف وتعلم أحـكام التجويد، نوجهه إلى بحث الدكتورين منصور الغامدي وعبدالله الأنصاري [١٣] المقدم في ندوة القرآن

الكـريم والتقنيات المعاصرة "تقنية المعلومات" بمجمع الملك فهد لطباعة المصحف الشريف سنة ٢٠٠٩، وكذلك بحث الدكتور غانم قدوري [١٤] وأحمد راغب [١٥].

تضم هذه الورقة – غير المقدمة والخاتمة – قسمين أساسيين: قسمًا نستعرض فيــه بنوع من التفصيل الخلفية العلمية لتقنيات التعـرف الآلي على الكلام وأدواته ومنهجيــة بنــاء تطبيقاته، في حين نقدم في القسم الآخــر أهم الأعمال التي اهتمت باستخدام تقنيات التعرف الآلي على الكلام في مجال القرآن الكريم كرصد للواقع الفعلي لاستغلال هذه التقنيات في خدمة القرآن.

تقنيات التعرف الآلي على الكلام وأدواتها

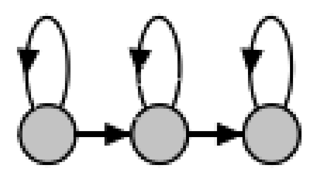
لقــد أفضت الجهود البحثية في مجال التعرف الآلي على الكلام إلى نجاحات معتــبرة وتجســدت بشكل كبــير في العقــود الثلاثة الأخــيرة مع استخــدام الطرق الإحصائية، وخصوصًا ما يعتمد منها على نماذج مركوف الخفية (Hidden Markov Models - HMM) [١٦–١٨]. وسنستعرضــ بتفصيــل، في الأجــزاء التاليـة، منهجيــةَ التعاطي مع تقنيات التعرف الآلي على الــكلام المعتمد على هذا النوع من النمذجة.

منهج بناء أنظمة التعرف الآلي على الكلام المعتمد على نماذج ماركوف الخفية

يعتمــد مبدأ استخدام نمــاذج مركوف الخفيــة (HMMs) [١٩] في التعرف الآلي علــى الكلام، علــى تقسيم مفردات المعجــم اللغوي المراد التعــرف عليها إلى سلسلــة مــن الوحدات الصوتيــة التي يتــم تمثيل كل واحــدة منها بنمــوذج صوتي (Acoustic Model) مكــون مــن مقاطع متواليــة (Sequence of States)، حيث يمثــل كل مقطــع في النمــوذج قسما متجانســا (او أجزاء متجانســة) من الإشارة الصوتية للصوت الممثل.

البنية النمطية لنماذج ماركوف الخفية في تمثيل الأصوات المنطوقة

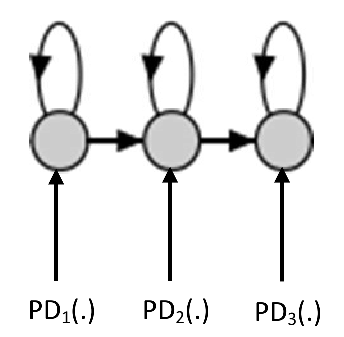
تبين أغلب الدراسات الحديثة أنه يكفي استخدام نماذج مركوف خفية ذوات ثلاث مقاطع رئيسية (Three Emitting States) لتمثيل الأصوات في بنيتها الأساسية]١[، بحيث يمثل المقطعان الأول والأخير الأجزاء الانتقالية مع الوحدات الصوتية المجاورة في حين يمثل المقطع الأوسط الجزء الثابت والمستقر من الصوت نفسه؛ الشكل ١ يقدم تمثيلا تصوريًا لهذا النموذج.



الشكل ١: البنية الأساسية لتمثيل الوحدات الصوتية

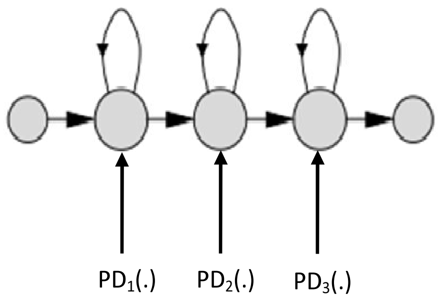
وفي هذا النموذج، يلاحظ أن الأجزاء المتشابهة (أو المتجانسة) في بداية الإشارة الصوتية يتم تثبيتها على المقطع الأول في النموذج من خلال السهم الذي يخرج من المقطع ويعود إليه؛ ومع التقدم الزمني في الإشارة الصوتية وتغير خصائصها يتم الإنتقال إلى المقطع الثانئ (الأوسط) الذي بدوره تثبت عليه الأجزاء المتشابهة من الإشارة الممثلة للقسم الثابت والمستقر من الصوت عبر السهم الذي يخرج من المقطع ويعود إليه؛ بعد ذلك يتم الإنتقال إلى المقطع الثالث لتثبت عليه الأجزاء المتبقية من الإشارة الصوتية كذلك بواسطة السهم الذي يخرج من المقطع ويعود إليه.

ولضبط عمليات مطابقة أجزاء الإشارة على مقاطع النموذج الثلاثة، يتم استخدام توزيعات احتمالية (Probability Distributions-PDs) ترفق بالمقاطع لتمثيل الجزء الصوتي المصاحب، (أنظر الشكل ٢).



الشكل ٢ : البنية الأساسية لتمثيل الوحدات الصوتية مع التوزيعات الإحتمالية

ومـن الناحيــة الإجرائية، يضاف عــادة إلى المقاطع الثلاثــة (المذكورة آنفا) مقطعــان استثنائيــان، واحد لتمثيـل البداية والآخـر لتمثيل النهايــة، وذلك لربط عناصر السلسلة الصوتية بعضها ببعض (انظر الشكل ٣).



الشكل ٣ : البنية الإجملية لتمثيل الوحدات الصوتية

وبناء على ما سبق فإن كل وحدة صوتية تمثل بنموذج مركوف خفي (HMM) (انظر الشكل ٤) يرمز له بـ $\lambda = (\prod, A, B)$ ،

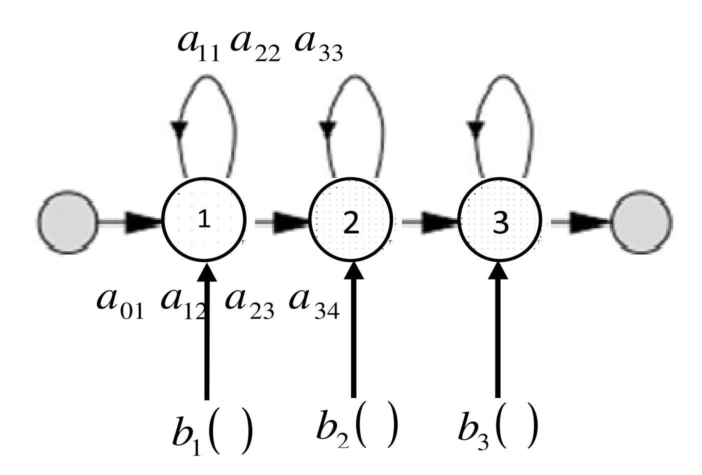
حيث:

أ. $\prod = (\pi_i)_{1 \le i \le N}$: تمثل توزيع القيم الاحتمــالات الأولية لمقاطع النموذج، وهــي دائمًا تســاوي "واحد" عنــد المقطع الذي يمثـل مدخل النمــوذج ومنطلقه ($\pi_1 = 1$)، و"صفر" لبقية المقاطع ($\pi_i = 0, i = 2, \ldots, N$)؛ أما N فترمز لعدد المقاطع في النموذج.

ب. $A = (a_{ij})_{1 \le i,j \le N}$: مصفوفـة تضم قيمًا عددية تمثـل احتمالات الانتقال الممكنـة والمسموح بها بين مقاطع النمـوذج، وتحدد الهيكل أو البنية العامة للنموذج (Model Topology).

ت. $B = (b_i(.))_{1 \le i \le N}$: مجموعـة التوزيعـات الاحتماليـة المرفقـة بمقاطـع النمـوذج؛ حيث يرمز $b_i(.)$ للتوزيع الاحتمـالي المرفق بالمقطع رقم i في النموذج. ومن خلال هذه التوزيعات يتم التعرف على الأجزاء الصوتية أثناء مطابقة الإشارة على مقاطع النموذج.

نشـير إلى أن التوزيعـات الاحتماليـة $b_i(.)$ عادة ما تكون من نـوع (Gaussian)؛ وقـد تكون هذه التوزيعات منفردة بواقع واحـد لكل مقطع (Single Gaussian) بالنسبـة للتطبيقات البسيطة، أو خليط من التوزيعات (Gaussian Mixture Models - GMMs) في كل مقطـع من مقاطع النمـوذج بالنسبة للتطبيقات الأكثر تعقيدا وذلك لزيادة القدرة على تمييز الخصائص الصوتية وضبطها بشكل أدق.

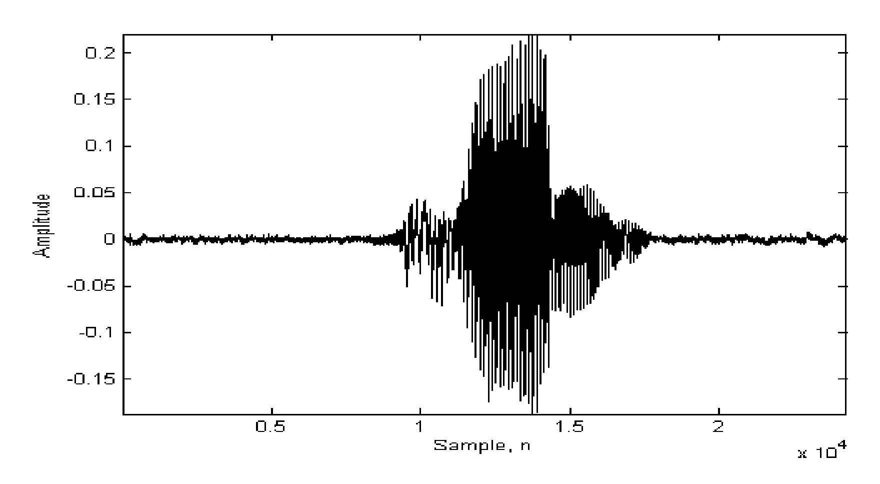


الشكل ٤ : بنية نمطية لنموذج ماركوف ذو ثلاث مقاطع أساسية

وتحتــاج نماذج ماركــوف الخفية بشكل عــام إلى عمليات تدريــب مكثفة على نسخ متعددة للوحدات الصوتية التي تتم نمذجتها لتتمكن التوزيعات الاحتمالية من تجميــع الخصائص الصوتية المختلفة حسب النطق. ومن أجل ذلك لا بد من وجود قاعدة بيانات صوتية معدة سلفًا وفقًا لمواصفات ومميزات مناسبة.

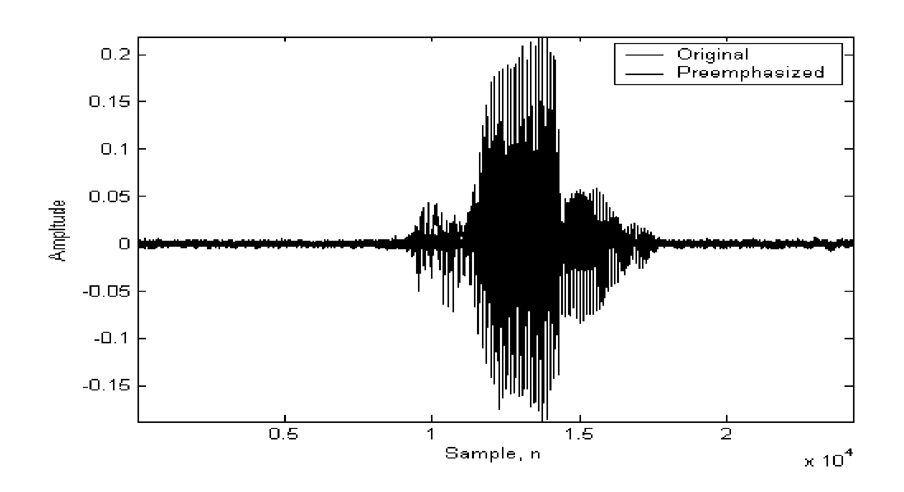
استخلاص الخصائص الصوتية من الإشارات لتدريب النماذج

يبدأ التحضير لعملية التدريب، بمعالجة الإشارات الصوتية في قاعدة البيانات واستخــراج الخصائــص منهــا (Feature Extraction)، وهــو مــا يقتضي تقسيم الإشارة إلى مقاطع أو نوافذ زمنية قصيرة (تتراوح عادة بين ١٠ و٢٠ ملسكند) يتم تمثيــل كل منها بمتجهات تضم مجموعة من الخصائــص المميزة للمقطع. وهناك أنواع من الخصائص وطرق متعددة لاستخلاصها، لكن أكثرها استخدامًا وشيوعًا في مجــال التعــرف الآلي على الكلام هو مــا يعرف بـــ MFCC (Mel Frequency Cepstral Coefficients). وللتحضــير لاستخلاصــ هــذه الخصائــص، لا بد من تحويل الإشارة من صيغتها العادية على شكل موجة إلى صيغة رقمية قابلة للتخزين والمعالجة (أنظر الشكل ٥). وهنا ينبغي تحديد معدل العينات (Sampling Rate) التــي تقيس عدد العينات المأخــوذة في الثانية من الإشارة الصوتية إضافة إلى سعة تمثيلهــا (عــدد البتــات المستخدمة). وبما أن أغلــب المعلومات في الــكلام البشري موجــودة في ترددات أقــل من عشرة آلاف هرتز، فإن عشريــن ألف عينة في الثانية كافيــة للتعرف على الكلام. فالــكلام المسجل مثلا بواسطة الهاتف لا يحتاج إلا إلى أربعة آلاف هرتز، أي أن ثمانية آلاف عينة في الثانية كافية له.



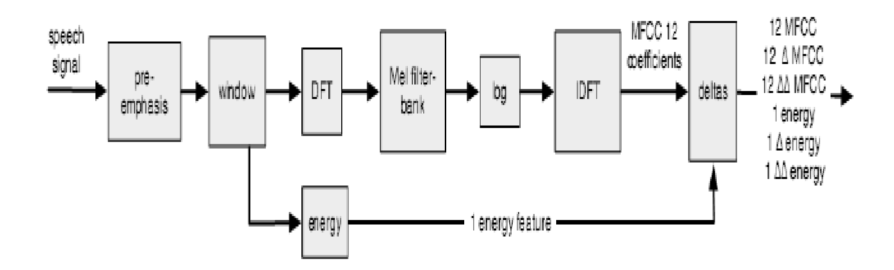
الشكل ٥ : إشارة صوتية مقسمة إلى عينات

وتمـر عملية استخلاص الخصائص أو المؤثرات الصوتيـة بعدة مراحل كما يبينها الشكل ٧. وتبدأ هذه المراحل بمعالجة الترددات العالية والمنخفضة لتكون في مستوى متناسب (preemphasis) (الشكل ٦).



الشكل ٦ : معالجة الترددات العالية والمنخفضة

يتم بعد ذلك تقسيم الإشارة إلى نوافذ زمنية صغيرة يتم تمثيل كل جزء منها بمتجـه يتكـون من مجموعة من القيم (عـادة ما تكون ٣٩ قيمة) تمثـل الإثنا عشر عنصـرًا الأولى من قيم MFCC مضافًا إليهـا قيمة الطاقة للنافذة الزمنية لتمثيل الخصائص الثابتـة في المقطع الصوتي، ثم يضاف إلى ذلـك قيمة الاشتقاق الأول والثاني لهـذه العناصر لتمثيل الخصائص الديناميكيـة للمقطع. نشير هنا إلى أن عـدد القيـم في المتجه الذي يمثل المقطع الصوتي يمكـن أن يكون أقل أو أكثر حسب أنـواع التطبيقات، إلا إن القيم المذكـورة هنا (٣٩ قيمة) تعتبر متوسطة وهي كثيرة الاستخدام في تطبيقات التعرف الآلي على الكلام.



الشكل ٧: خطوات استخلاص الخصائص الصوتية (مستقاه من [١])

تدريب النماذج الصوتية

يتم التدريب بإدخال الخصائص الصوتية المستخرجة من الإشارات الصوتية إلى النماذج بشكل تكراري حتى تصل النماذج إلى القيم المثلى (Optimal Values). وهـذا يعني أننا نقوم بتغيير قيم النمـوذج الصوتي التي رأيناها سابقا، أي مجموع قيم مصفوفة الانتقال بين مقاطع النموذج $(a_{ij})_{1\le i,j\le N}$ وقيم التوزيعات الاحتمالية المرفقـة بمقاطع النموذج $(b_i(.))_{1\le i\le N}$. يتم ذلك من خلال قاعدة بيانات التدريب التـي يفترض أن تضم مجموعـة من التكرارات (نطق متكرر) لـكل وحدة صوتية. ويمكـن أن يتـم التدريـب بطريقتين، تدريب منفصـل للوحـدات الصوتية يعرف بـ (Isolated-Unit Training) أو تدريب مدمج (Embedded-Unit Training).

فالنوع الأول يقتضي أن يكون كل ملف صوتي في قاعدة بيانات التدريب مقسمًا ومرمـزًا بحيـث تحدد البداية والنهاية الزمنية لكل وحـدة صوتية فيه؛ ويتم تجميع الأجـزاء الخاصة بكل وحدة صوتية من مختلف الملفات وتستخدم مع بعض لتدريب النمـوذج الذي يمثل تلـك الوحدة الصوتيـة. وتعتمد جودة التدريـب هنا على دقة تحديـد الحـدود الزمنية للأصوات. ونظرًا لما يتطلبه ذلـك من وقت وجهد فإنه قد يكـون من الصعب تجهيز قواعد بيانات صوتية كبـيرة وفقا لهذه الطريقة وبالتالي يتـم اللجوء إلى الطريقة الثانيـة التي تعرف بالتدريب المدمـج. وفيها يتم الاكتفاء بتحديـد قائمـة الوحدات الصوتية لـكل ملف صوتي ثم يـترك للخوارزميات مهمة تقسيـم الإشـارة وتحديد الحـدود الزمنية بشـكل آلي من خلال مطابقـة الإشارة الصوتية على تلك القائمة. ومع التقدم في مراحل التدريب تعمل الخوارزميات على ضبط تلك الحدود بشكل أفضل تدريجيًا حتى تصل إلى الأمثل. ويبقى وجود بيانات صوتيـة مقسمة زمنيًـا على مستوى الوحدات أمرًا بالـغ الأهمية في تسريع التدريب وضبطه بشكل أفضل، إضافة إلى أهميته في عدة أنواع من التطبيقات.

وقبـل الشروع في التدريب بأي من الطريقتين، يتم عـادة تحديد قيم ابتدائية للنمـوذج تكـون منطلقًا لـه؛ ويتم ذلك في الطريقـة الأولى بتقسيـم الإشارة بشكل متساو بين مقاطع النموذج ثم يتم استخدام إحدى صور الخوارزمية المعروفة بـاسم صاحبهـا Viterbi [٢٠-٢١] لتحديد التسلسل الأمثـل؛ وخوارزمية Viterbi تعتمد علـى مبدء البرمجة الديناميكية، وقـد ظهرت لأول مرة في منتصف الستينيات من القـرن الماضـي. أما في الطريقة الثانيـة فيتم حساب متوسـط وانحراف عام على كل بيانـات التدريب ويتم استخدامهما كقيـم ابتدائية لكل التوزيعات الاحتمالية في النمـوذج (وهذا النمط يعـرف بـ"الإنطلاقة المنبسطة أو المستويـة" Flat Start)؛ وبالنسبـة لاحتمـالات الانتقال بين مقاطـع النموذج فيكفي أن توضـع بشكل متساو في البدايـة على أن تتغير أثناء التدريب. بعـد تلك التحضيرات الأولية يتم التدريب الفعلي باستخـدام خوارزمية تعرف هي الأخرى بأسماء أصحابها Baum-Weltch

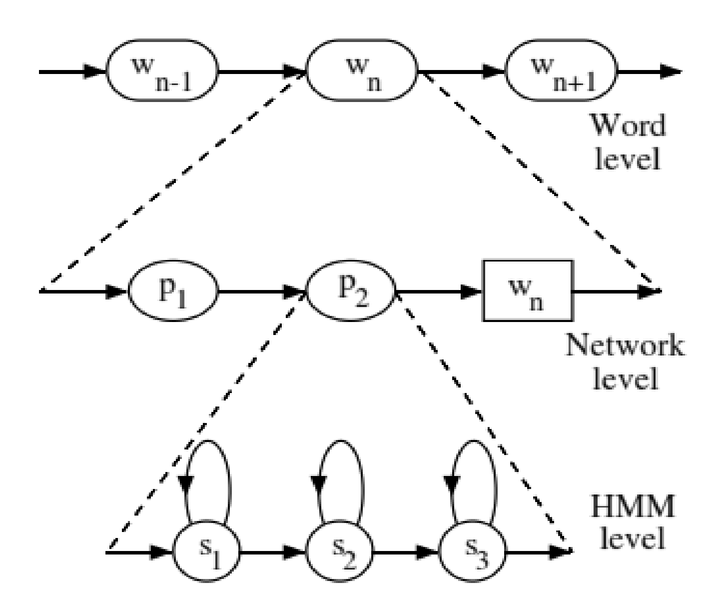
[٢٢]، وتعمل على حساب القيم المثلى لنماذج ماركوف الخفية بإدخال الخصائص الصوتية بشكل تكراري.

اختبار وتجريب النماذج

بعـد الانتهاء من تدريـب النماذج الصوتية، يتم عـادة العمل على قياس مدى نجاعتهـا من خلال عرض عينات صوتية عليها ومحاولـة تحديد ما نُطِق ثم قياس مـدى صحته. ويحتاج هذا الأمر إلى وضع نموذج لغوي (language model) يحدد التتابعـات المسموح بها حسب نوع التطبيـق؛ فقد يكون التطبيق عبارة عن كلمات أو وحدات منفردة يتم نطق كل منها على حدة والتعرف عليه بشكل منفصل (ويعرف هذا النـوع بـ Isolated-Words Recognition أي التعرف علـى كلمات منفردة)، ويكـون النمـوذج اللغوي في هذه الحالة بسيطا؛ وقد يكـون التطبيق عبارة عن جمل تضـم سلسلة من الكلمات وفقا لقواعد يتم تحديدها وتشكل طبيعة النموذج اللغوي (ويعرف هذا النوع بـ Connected–Words Recognition أي التعرف على كلمات متصلـة أو مترابطـة)؛ وقد يكون التطبيق يسعى إلى التعـرف الآلي على الكلام في صيغته الطبيعية (وهو ما يعرف بـ Continuous Speech Recognition أي التعرف علـى الكلام المتواصل)، ويبنى النموذج اللغوي في هذه الحالة وفقًا للقواعد اللغوية التـي تحكم بنيـة الجملة. ويستخدم الجانـب النصي في ذخـيرة التدريب لحساب احتمـالات التتابعات الممكنة وفقًا لنماذج إحصائيـة تعرف ب"N-Gram". ويلعب النموذج اللغوي دورًا مهمًا في تحسين نسبة التعرف لأنه يساعد على تحديد الاختيار الأمثل من بين الاحتمالات الممكنة بناء على التسلسلات التي تدرب عليها.

ففي مرحلة التعـرف يتم أولا بناء فضاء للبحث (Search Graph) على شكل شبكـة تربـط بين الوحـدات الصوتية وفقًـا لما يحـدده النموذج اللغوي؛ فـإذا كان النمـوذج اللغوي معرفًا على مستوى الجمل، فإن الشبكة تبنى على ثلاثة مستويات، الأول يمثـل كيفية توالي الكلمـات في الجملة، أما الثاني ففيه يتـم تعويض كل كلمة

بحالاتهـا النطقية وفقًا لما يحدده القاموس النطقي (Pronunciation Lexicon)، وفي المسـتـوى الثالث يتم تعويض كل وحـدة في القاموس النطقي بالنموذج الصوتي الـذي يمثلهـا (انظر الشـكل ٨). أما إن كان النمـوذج اللغوي معرفًـا على مستوى الوحـدات اللغوية الأصغر من الكلمـة، فيتم تعويض كل وحـدة بنموذجها الصـوتي مباشرة للحصول على الشبكة.



الشكل ٨: توليد فضاء أو شبكة البحث في مرحلة التعرف

يتم البحث في هذه الشبكة – أيضًا – باستخدام خوارزمية البرمجة الديناميكية Viterbi التـي تعمل على إيجاد أفضل مسار في الشبكـة، وهو عبارة عن سلسلة من الوحـدات اللغوية، حيث يتم تحويلها بواسطة القاموس النطقي إلى كلمات ثم إلى جمل عبر النموذج اللغوي؛ وذلك في مسار عكسي لبناء الشبكة.

قياس أداء أنظمة التعرف

لقياس أداء نظام التعرف، يتم مقارنة ما تم التعرف عليه بواسطة خوارزمية

Viterbi مـع مـا تم نطقـه أصلا، حيـث ينبغي أن تكـون هناك عينة مجهـزة سلفًا وموضوعـة في صيغة مماثلـة لما يخرجه نظـام التعرف؛ وتعرف هـذه العينة بعينة الاختبـار، وهي جزء من قاعدة البيانـات الصوتية التي أشرنا إلى ضرورة وجودها خلال الحديث عن التدريب.

وكوحـدة لقياس الأداء، يتم حساب عدد الكلمات أو الوحـدات الصوتية التي تم التعرف عليها بشكل صحيح (دون النظر إلى مكانها وترتيبها حسب المدخلات) وتقسـم على عدد الكلمات المدخلة؛ ويعرف هذا المقياس بـ "Word Recognition Rate" أي نسبـة التعـرف على الكلمـات. ويوجد مقياس أكثر صرامـة لقياس دقة نظام التعرف (يسمى "Recognition Accuracy" أي دقة التعرف)، حيث تحتسب فيه نسب الاستبـدال والإضافة والحذف، بمعنا أن الكلمات التي يتم التعرف عليها بشكل صحيح لا تحسب إلا إذا كانت في مكانها المناسب حسب ترتيب المدخلات.

أدوات بناء أنظمة التعرف الآلي على الكلام

يعتـبر بنـاء أنظمـة التعرف الآلي علـى الكلام بشكل عـام أمـرًا صعبًا لكونه يحتاج إلى جهـود كبيرة وتخصصات مختلفة (صوتية ولغوية وحاسوبية وغيرها). ولتسهيـل وتسريـع بناء تلك الأنظمـة تم التوجه في الأوساط الأكاديميـة إلى إنشاء بيئات تطويرية توفر جملة من الأدوات الأساسية يمكن للمطور الاستعانة بها في بناء تطبيقاتـه المختلفة. ولعل من أشهر تلـك البيئات وأكثرها استخدامًا نظام Sphinx [٢٣] وحزمـة أدوات HTK (Hidden Markov Model Toolkit) [٢٤] اللـذان يمكنان من بناء أنظمة التعرف الآلي على الكلام بمختلف أنواعها، وذلك من حيث:

١. الحجـم، أي عدد الوحدات التي يمكن للنظـام التعرف عليها "vocabulary"؛ ويمكن أن يكون هذا العدد صغيرا أو متوسطا أو كبيرا،

٢. طريقـة النطق، أي كيفية التعاطي مـع المدخلات في النظام؛ هل يتعامل النظام مع كلمات أو وحدات منفصلة، أم يتعامل مع كلمات متصلة تشكل جملا قصيرة مبنية وفقا لقواعد محددة، أم يتعاطى مع الكلام الطبيعي في مجمله،

٣. عدد المستخدمين، أي عدد المتحدثين الذين يمكن للنظام التعرف عليهم. فقد يكون للنظام مستخدما واحدا تدرب على صوته ولا يتعرف إلا عليه، ويدعى النظام في هذه الحالة "speaker dependent"؛ وقد يكون النظام قادرا على التعرف على عدد معين من المستخدمين تدرب على أصواتهم، ويدعى في هذه الحالة "multi-speakers"؛ وقد يكون النظام مستقلا عن المتحدث، بمعنى أنه قادر على التعرف على صوت أي متحدث لكونه تدرب على عد كبير من المتحدثين، ويدعى في هذه الحالة "speaker in-dependent".

تم تطوير نظام Sphinx في جامعة كارنيجي ميلون بالولايات المتحدة الأمريكية، وهو يستخدم نماذج ماركوف الخفية في نمذجة الأصوات ويضم قسمين رئيسين، قسم للتدريب وآخر للتعرف. وقد ظهرت منه إصدارات مختلفة تستخدم طرق مختلفة في التعرف آخرها Sphinx4 [٢٥] الذي تم تطويره باستخدام لغة الجافا وهو يحتوي على ثلاثة أجزاء أساسية أساسية هي:

١. الواجهة (FrontEnd): تقوم باستقبال الإشارة وتحويلها إلى مجموعة من الخصائص الرقمية وفقًا لآليات استخلاص الخصائص التي شرحت في الجزء السابق.
٢. اللغوي (Linguist): يقوم بترجمة النموذج اللغوي للنظام ودمجه مع معلومات القاموس اللفظي الذي يضم الحالات النطقية الممكنة والنماذج الصوتية لتكوين مجال وشبكة البحث.
٣. المتعرف (Decoder): يقوم باستخدام الخصائص الرقمية ومجال البحث الذي تم إنشاؤه لأداء عملية التعرف الفعلية.

أما حزمة أدوات HTK التي تعتبر الأشهر والأكثر انتشارا، فقد تم تطويرها في جامعة كامبريدج ببريطانيا. وفيها يتم تمثيل كل وحدة صوتية باستخدام نموذج

ماركوف خفي واحد، وكل كلمة كسلسلة من النماذج. ويتكون HTK من أربعة أجزاء رئيسية هي:

١. أدوات تجهيـز البيانـات: وتقـوم باستقبال الإشـارة الصوتية ومعالجتها ثـم استخلاص الخصائصـص الرقمية منها وفقًا لأحـدى الصيغ المناسبة (MFCC، LPC، إلخ.)؛ كما تقوم هذه الأدوات بتجهيز البيانات للمرحلة المقبلة.
٢. أدوات التدريـب: وتمكـن مـن تعريـف الهيكل العـام للنمـاذج الصوتية وتحديـد قيمها الابتدائية ثم تدريبها بالآلية والخوارزميات التي ذكرناها في الجزء السابق.
٣. أدوات التعـرف: وتمكن من استخدام النماذج الصوتية والنموذج اللغوي والقاموس النطقـي مع خوارزميـة Viterbi لمحاولة التعـرف على ما تم نطقه.
٤. أدوات التحليل: تستخدم لتحليل النتائج وتقييم الأداء.

ولعل هذين النظامين يعتبران من أنضج الوسائل المعتمدة على الطرق الإحصائية والمستخدمـة اليـوم في بناء أنظمـة التعرف الآلي على الـكلام. ولا توجد حسب علمنا دراسات مقارنة تبين أهمية واحد على الآخر حسب نوع التطبيق المراد عمله.

واقع استخدام تقنيات التعرف الآلي على الكلام في خدمة القرآن الكريم

إن الأعمـال البحثية في مجال التعرف الآلي علـى الكلام العربي بشكل عام لا تزال محدودة إذا ما قورنت بما تم في اللغات الأخرى كالإنكليزية مثلا، على الرغم مـن أن اللغة العربية تصنـف من حيث عدد المتحدثين بها في المرتبـة الثانية عالميًا بأكثر من ٤٢٢ مليون متحدث في حين تأتي اللغة الإنكليزية في المرتبة الرابعة وذلك وفقـا لمـا أوردته موسوعة أنكرتا الشهيرة وما جاء في هـذه الإحصائية [٢٦]. ولعل من بين الأسباب التي قد تكون وراء ذلك النقص غياب البنى التحتية اللازمة لبناء

أنظمـة التعرف الآلي على الكلام، سواء ما تعلق منهـا بالصوتيات العربية وقواعد بياناتهـا [٢٧] أو تقنيات معالجتها كلغة طبيعيـة [٢٨] أو غير ذلك. ومع هذا فقد بـدأت تلوح بـوادر لتزايـد الاهتمـام في السنوات الأخـيرة، حيث انطلقـت أبحاث متعددة، نذكر من بينها [٢٧-٣٤].

أمـا الأعمال التي اهتمـت بتوظيف تقنيات التعـرف الآلـي على الكلام لخدمة القـرآن الكريم فهي قليلة علـى حد علمنا، وبعضها عبارة عن محاولات متواضعة لا تعكس أهمية القرآن الكريم في حياة المسلمين.

ومـن بين هذه الأعمال التي اهتمت بتوظيف تقنيات التعرف الآلي على الكلام لخدمـة القرآن الكـريم، ما قام به مرتجى وزمـلاؤه في جامعة عمـان، حيث سعوا لبنـاء نظام للتعرف الآلي على التلاوة القرآنيـة مرتبط بالمتحدث (Speaker-Dependent) واستخدمـوا في ذلك نماذج ماركـوف الخفية لتمثيل الوحدات الصوتية حسب السياق (tri-phone HMM model) [٣٥]. وقد استخدموا عينة من ٢٥ آية للتدريب والاختبار، ووصلت نسبة التعرف عندهم إلى ٨٠٪ حيث اعتبروها مناسبة ومرضيـة جـدا إذا مـا قورنت مع قلـة الآيات المستخدمـة. وقد قام نفسـ الفريق، لاحقا، بتطويـر النظام ليصبح مستقلا عن المتحـدث (Speaker-Independent) واستخدمـوا تقنيـة MLLR (Maximum Likelihood Linear Regression) لملاءمة النموذج [٣٦]. وقد أخذوا الجزء الأخير من القرآن الكريم (جزء "عم") بتـلاوات خمسة قراء للتدريب والاختبار ولكنهم لم يحددوا ما الحجم المستخدم في التدريـب والاختبـار كل على حدة. كما أنهم لم يبينوا مـا إذا كان الاختبار وقع على جـزء مما استخدم في التدريب أم لا؟ ولم يذكروا كذلك ما إذا كانوا قد استخدموا نموذجًـا لغويًا أم لا؟ مـع أن لكل ذلك علاقة وطيدة بـأداء النظام. وقد وصلوا إلى نتائـج تـتراوح بـين ٦٨٪ و٨٥٪ حسب القـارئ، أي بمعدل حـوالي ٧٧٪. وأوردوا في بحثهـم أن هذه النتائج تم الحصول عليها باستخدام نماذج ماركوف مكونة من ١٣

مقطـع لكل واحد على حدة، ونظرًا لذلك فقد ذكروا أن مرحلة التعرف في نظامهم تعتبر بطيئة. واستخدموا في بناء تلك الأنظمة أدوات HTK.

وقـد أهتم طبال وزملاؤه في الجامعـة اللبنانية باكتشاف الآيـات القرآنية في الملفات الصوتية بغية تحديدها واستخلاصها [٣٧]. وقد استخدموا نظام Sphinx في ذلك، ولكنهم اقتصـروا فقط على سورة الإخلاص بمقرئين متعددين (عشرين قارئـا) ولم يبينـوا ما إذا كان الاختبـار تم على نفس التـلاوات التي استخدمت في التدريـب. وعلـى اعتبار أنهم يسعون فقـط إلى تمييز الآيـات القرآنية عن غيرها، فقـد استخدموا نموذجا لغويًا للآيـات الموجودة في سورة الإخلاص. ويذكرون أنهم قاموا بتجارب على نوعين من التلاوات، تلاوات مرتلة وتلاوات مجودة، وحصلوا في المعدل على نسبة ٩٢٪ في النوع الأول و٩٠٪ في النوع الثاني مع العلم أنهم استخدموا ٢٠ قارئـا في كل نـوع. وحسبمـا أوردوا فإن التلاوات المرتلة أسـرع من تلك المجودة وأقل تركيزا منها على استيفاء المدد الزمنية للأصوات القرآنية.

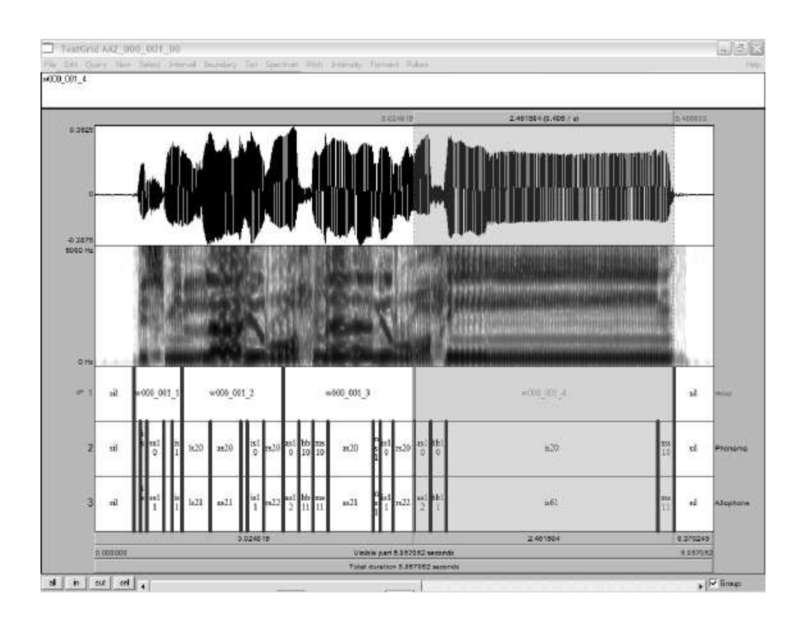
ومـن بين الأعمـال الهامة في المجال، مشـروع "J-QAF" [٣٨] الذي أطلقته الحكومـة الماليزية سنـة ٢٠٠٤م ضمن خطتها لتطوير الرأس المـال البشري "En-hancing Human Capital" ورصـدت لـه مبالغ مالية كبـيرة؛ وهو يسعى إلى بناء أنظمـة حاسوبيـة لتعليم تلاوة القرآن الكريم واللغة العربيـة والواجبات الدينية في المدارس الابتدائيـة إضافـة إلى تعلم الخط المـلاوي العربي. هـذا المشروع يتكون مـن سبع وحدات وأقسـام تعليميـة (Teaching Modules) بنيـت بشكل تدريجي حتـى نهايـة الفـترة الرئيسيـة للمشروع سنـة ٢٠١٠م؛ ومـن بين هـذه الوحدات، أقســام لتحفيظ القرآن الكـريم وتعليم قواعده التجويدية إضافـة إلى قسم لتعليم اللغـة العربيـة. والأقسام التي تعنى بالقرآن الكريم في هـذا المشروع بدأت بتقنيات الملتيميديـا التفاعليـة [٣٩-٤٠] ثم تدرجت لتصل في المراحـل الأخيرة إلى إدخال تقنيات التعرف الآلي على الكلام حسب ما ذكر [٤١]؛ ولم يتسنَّ لنا إيجاد تفاصيل

حـول آليــات التعرف الآلي المستخدمــة ولا مستوى الدقة الذي وصلـت له. وقد قام ظريـف وزملاؤه مؤخرًا بدراسة لتقييم فاعلية أقسام تعليم القرآن في هذا المشروع وقياس أثرها على الطلاب [٤٢]، وأظهرت الدراسة مستوى عاليا في ضبط التلاوة القرآنية (النطق وأحكام التجويد) وانعكاسات ذلك على تحسن مستويات الطلاب في تعلم اللغة العربية.

ومـن الأعمــال البارزة في مجـال توظيف تقنيـات التعـرف الآلي على الكلام لخدمـة القـرآن الكريم، مشـروع حفص [٤٣] الذي أنتجتـه الشركـة الهندسية لتطويـر نظـم الحاسبات المصرية RDI في صـورة برنامج كمبيوتـر تفاعلي يسعى للتمكين مـن التعلُّم الذاتي لأحكام تجويد القرآن الكريم. تقوم الفكرة على محاولة البرنامج محاكاة المقرئ في الكتاتيب من خلال إسماع المتعلم مثالًا لتلاوة نموذجية ثـم يَطْلُبُ منــه أن يحاكيها قدر استطاعته؛ يقوم البرنامــج بتحليل ما نطقه المتعلم وتوليد تقرير صوتي وكتابي بأخطاء المتعلم ويطلب منه التركيز على تصحيح نطقه للمواضيـع التي أخطأ فيها ثم إعــادة التسميع؛ وهكذا إلى أن يجيز البرنامج قراءة المتعلم. وقد نشرت أوراق بحثية تبين الجوانب التقنية لآلية التعرف الآلي المتبعة في تحديـد الأخطاء النطقية خلال تــلاوة المتعلم للأمثلة المضمنـة في البرنامج والتي تــدور حول مواطن وجـود أحكام تجويدية [٤٤-٤٦]. وقد ذكـر أن تقييم البرنامج يتم من خلال قاعدة بيانات تضم تلاوات للأمثلة التي يتعامل معها البرنامج لقراء مجازيـن؛ حيث قـرئ كل مثال عدة مـرات، بعضها بشكل صحيـح وبعضها أدخلت فيــه أخطاء نطقية شائعة. ثم طلب من خبراء لغويين تحليل مختلف النطق المسجل وكتابتــه صوتيــا مع توصيــف للأخطاء ليستخــدم كل ذلك في آليــات التعرف الآلي والمقارنــات أثنــاء استخدام البرنامج. وتشير النتائــج إلى أن البرناج يعطي القرار الصحيـح حول القــراءة (صحة القــراءة أو رسالــة تشخيص اللأخطــاء) بنسبة ٨٤٪، وهــي نسبة جيدة. وتذكر الشركة أنها تعمل على تحسين أداء البرنامج بشكل

مستمر، وتسعى لتطويره مستقبلا ليتجاوز حدود التعامل مع حالات نطقية محددة ويصبح قادرًا على تصحيح التلاوة والاستخدام في التحفيظ الآلي بشكل عام.

ومن ضمن الأعمال البارزة في المجال أيضا، مشروع التعليم الآلي للقرآن الكريم [٤٧] الذي مولته مدينة الملك عبدالعزيز للعلوم والتقنية، والذي سعى إلى وضع نموذج لبيئة حاسوبية للتحفيظ الآلي للقرآن الكريم تستخدم فيها تقنيات التعرف الآلي على الكلام لمساعدة القارئ على تصحيح نطقه وضبط حفظه. وقد ركز المشروع على إعداد بنى تحتية قوية تساعد على الوصول إلى مصحح آلي دقيق يمكن استخدامه في نظام حلقات افتراضية لتعليم القرآن الكريم. ومن أهم البنى التحتية التي ركز عليها المشروع، بناء قاعدة بيانات صوتية لتلاوات قرآنية وتجهيزها بشكل مناسب لبناء أنظمة للتعرف الآلي على الأصوات القرآنية بمختلف صورها وكذلك أصوات اللغة العربية الفصحى حيث هي لغة القرآن. وقد استخدم جزء "عم" كعينة وتم تسجيله بتلاوات عشرة قراء تم اختيارهم من طلاب الجمعية الخيرية لتحفيظ القرآن الكريم بالرياض؛ وتمت التلاوات والتسجيلات تحت إشراف مختص في القرآن وعلومه. هذه التلاوات تم تقطيعها وترميزها صوتيا على ثلاثة مستويات هي الكلمة والأصوات الأصول (الفونيمات) ثم الأصوات الفروع (الآلوفونات) التي تمثل مختلف الحالات النطقية مثل الغنة والقلقلة والتفخيم والترقيق ودرجات المدود وغير ذلك من الظواهر الصوتية في التلاوة القرآنية [٤٨-٤٩]؛ والشكل ٩ يوضح مستويات التقطيع وآلية الترميز المستخدمة.

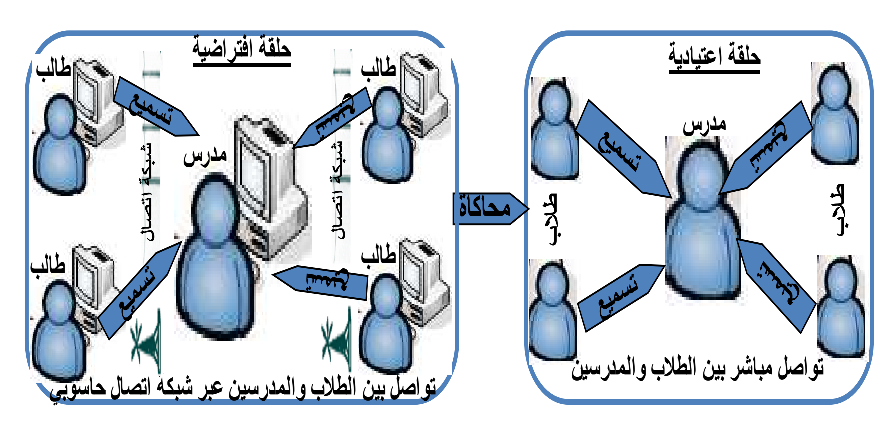


الشكل ٩ : مستويات التقطيع في قاعدة البيانات الصوتية للقرآن الكريم

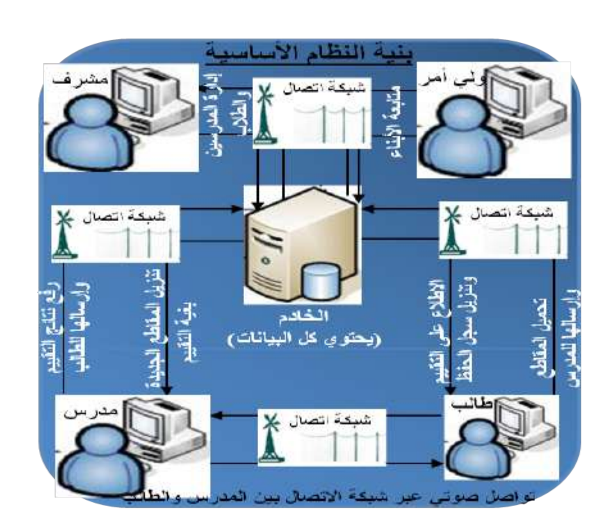
بالاعتماد على هذه القاعدة الصوتية تم بناء نموذجين أوليين للتعرف الآلي على الأصوات، واحد على مستوى الأصوات الأصول [٥٠] والآخر على مستوى الأصوات الفروع [٥١]. وقد استخدمت نماذج ماركوف الخفية مع خليط من التوزيعات الاحتمالية؛ وتم قياس الأداء من خلال سلسلة من التجارب، تم في كل واحدة منها الاختبار على عينة من قارئ محدد لم تستخدم في التدريب الذي يتم على بقية تلاوة هذا القارئ وتلاوات القراء الآخرين، ثم يحسب معدل نسب التعرف الإجمالية لكل القراء. وقد وصلت النتائج إلى ٩٢٪ بالنسبة للأصوات الأصول و٨٨٪ بالنسبة للأصوات الفروع؛ وهي نسب جيدة وتوحي بإمكانية الوصول إلى مستويات عالية من الدقة في التعرف على مختلف الأصوات القرآنية أثناء التلاوة.

وفي المشروع تم وضع تصور لنظام حلقات إفتراضية [٥٢-٥٣] تحاكي نظام التحفيظ الاعتيادي (الأشكال ١٠ و١١)، وتسهل التواصل بين المعلمين والمتعلمين

خصوصًـا لمن هـم في بلدان غير إسلاميـة. وقد وضعت آلية لربـط المصحح الآلي المعتمـد علـى تقنيـات التعرف الآلي علـى الكلام مع نظـام الحلقـات الافتراضية لمساعـدة المتعلم علـى تصحيح نطقه وتجريب حفظه قبل عرضـه على المدرس من خلال آليات التواصل.



شكل ١٠ : محاكاه حلقات التحفيظ (مستقاه من ٣٥)



شكل ١١ : معمارية نظام الحلقات الافتراضية (مستقاه من [٥٣])

وقد اهتم المشروع في جانب آخر من جوانبه بدراسة المتشابه اللفظي وتحديد مواطنه لأهمية ذلك في التحفيظ. وقد استخدمت تقنيات في مجال تحليل النصوص لقياس درجة التشابه بين الكلمات ووضع معايير ومقاييس لترتيب الآيات حسب مستويات التشابه [٥٤]. ووضعت آلية لربط التشابه مع التحفيظ لمساعدة الدارس في إتقان حفظه. كما أضيف في نظام التحفيظ مكتبة تضم أهم كتب علوم القرآن الأخرى (الإعراب، التفسير، أسباب النزول) وتم ربطها مع النص القرآني لتسهيل الوصول إليها.

وبشكل عام، فإن مشروع التعليم الآلي للقرآن الكريم وضع أسسًا قوية لبناء بيئة حاسوبية متكاملة للتحفيظ الآلي للقرآن الكريم، لكنه يحتاج إلى جهة تتبناه وتحوله من نموذج بحثي إلى منتج فعلي يخدم الراغبين في حفظ القرآن الكريم وتعلم علومه.

الخاتمة

لقـد قدمنـا في هذا البحـث بسطًا لأشهـر تقنيات التعـرف الآلي على الكلام ومنهجيـة بناء نظمـه وأدواته؛ ثـم استعرضنا بنوع مـن التفصيل واقـع استخدام هذه التقنيـات في الوقت الحالي لخدمة القرآن الكريم. ولعـل القارئ والمتتبع لهذه الأعمـال يشعـر بأنها لا تـزال محدودة وتحتـاج إلى تراكم في الخـبرة وتضافر في الجهـود لنصـل إلى مستـوى من التحكم في تلـك التقنيات يمكن مـن تطويعها بما يتماشى مع خصوصية القرآن الكريم وقدسيته.

ومـن أهم وأنضج مـا استعرضناه من أعمال مشروع حفص ومشروع التعليم الآلي للقـرآن الكريم؛ فالأول منتـج تجاري والثاني نموذج بحثـي. فمشروع حفص يركـز على تعليم وتطبيق أحكام التجويد من خـلال التعامل مع مقاطع محددة من القـرآن الكريم، حيث يسمع المتعلم تلاوتهـا بقراءة نموذجية ثم يطلب منه محاولة محاكاتهـا ومن ثم يقوم بتوجيهه بناء على مستوى نطقه؛ وهو عمل جاد ومفيد. أما مشـروع التعليم الآلي للقرآن الكريم فله أهداف أوسع وأشمل، تضم في صلبها بيئة حاسوبيـة للتحفيظ الآلي للقرآن الكريم مشتملة على مختلف علومه. وقد وضع فيه نموذج لحلقات افتراضية تحاكي نظام حلقات التحفيظ الاعتيادية وتسهل التواصل مع المقرئين. وتم وضع تصور لربطه مع نظام التصحيح الآلي للنطق والحفظ ليكون ذلـك عونا للدارس من جهة، ومساعدا للمقرئ مـن جهة أخرى في تسريع التعاطي مع المتعلمين. والنظام يعتمـد في رؤيته على عدم إلزام المتعلمين والمعلمين بالتزامن أثنـاء العمل مع إتاحة خيار للتواصل المباشـر سواء لتقديم توجيهات وتعليقات عند الحاجة أو لإجراء اختبارات لمنح إجازات أو كذلك للقيام بمسابقات قرآنية.

المراجع

1. D. Jurafsky, J. H. Martin. (2008). Speech and Language Processing: An Introduction to Speech Recognition, Computational Linguistics and Natural Language Processing. Prentice Hall, ISBN: 10: 0131873210.

2. http://www.cs.cmu.edu/~listen/

3. J. E. Beck, K. Chang, J. Mostow, A. Corbett. (2005). Using a student model to improve a computer tutor's speech recognition. Proceedings of the AIED 05 Workshop on Student Modeling for Language Tutors, 12th International Conference on Artificial Intelligence in Education, Amsterdam, 18-22 July.

4. http://www.colit.org/

5. R. Col, S. Vuuren, B. Pellom, et. Al. (2003). Perceptive Anie mated Interfaces: First Steps Towards a New Paradigm for Human Computer Interaction. Proceedings of the IEEE, vol.91, no9, pp. 1391-1405.

6. http://www.esat.kuleuven.be/psi/spraak/projects/SPACE/

7. W. Mattheyses, L. Latacz, W. Verhelst. (2009). On the importance of audiovisual coherence for the perceived quality of synthesized visual speech. EURASIP Journal on Audio, Speech, and Music Processing.

8. http://www.readingcompanion.org/
9. http://www.scilearn.com/products/reading-assistant

١٠. غانم قـدوري الحمد. (١٤٠٦هـ). الدراسات الصوتية عند علمـاء التجويد. مطبعة الخلود. بغداد.

١١. أبـو بشر عمرو بن عثمان "سيبويه". (١٩٧٥م). الكتاب. تحقيق: عبد السلام محمد هارون، دار الكتاب العربي. القاهرة.

١٢. محمـد بـن شحادة الغـول. (١٤١١هـ). بغية عبـاد الرحمـن لتحقيق تجويد القرآن. دار بن القيم. الدمام.

١٣. منصور الغامدي، عبدالله الأنصاري. (٢٠٠٩م). التقنيات المعاصرة في خدمة القرآن الكريم. نـدوة القرآن الكريم والتقنيات المعاصـرة "تقنية المعلومات"، المدينة المنورة –المملكة العربية السعودية، ١٣-١٥ أكتوبر.

١٤. غانم قدوري الحمد. (٢٠٠٩م). استخدام صورة آلة النطق ومخارج الحروف في تعليـم قواعـد التلاوة "تأصيـل وتحليل". نـدوة القرآن الكـريم والتقنيات المعاصـرة "تقنية المعلومات"، المدينة المنورة –المملكة العربية السعودية، ١٣-١٥ أكتوبر.

١٥. أحمـد راغب أحمـد. (٢٠٠٤م). فنولوجيا القرآن: دراسة لأحـكام التجويد في ضوء علم الأصوات الحديث. أطروحة لنيل درجة الماجستير. جامعة عين شمس.

16. L. Rabiner, B.H. Juang. (1993). Fundamentals of Speech Recognition. Prentice Hall.

17. F. Jelinek. (1998). Statistical Methods for Speech Recognition. Cambridge, MA: MIT Press.

18. X. Huang، A. Acero، H. Hon. (2001). Spoken Language Processing، Prentice Hall PTR.

19. L. Rabiner. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE، 77(2).

20. T. K. Vintsyuk. (1968). Speech discrimination by dynamic programming. Cybernetics، 4(1)، 52–57. Russian Kibernetika، 4(1)، pp. 81-88.

21. A. J. Viterbi. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory 13 (2): 260–269. doi:10.1109/TIT.1967.1054010، April.

22. F. Jelinek، L. Bahl، R. Mercer. (1975). Design of a linguistic statistical decoder for the recognition of continuous speech. IEEE Transactions of Information Theory 21 (3): 250–6. doi:10.1109/tit.1975.1055384، May.

23. http://cmusphinx.sourceforge.net/

24. http://htk.eng.cam.ac.uk/

25. http://cmusphinx.sourceforge.net/wiki/tutorialsphinx4

26. http://www.scribd.com/doc/14436546/Languages-Spoken-by-More-Than-10-Million-People.

27. M. Algamdi. (2003). KACST Arabic Phonetics Database. The Fifteenth International Congress of Phonetics Science، Barcelona، 3109-3112.

28. M. Elshafei، H. Al-Muhtaseb، M. Alghamdi، (2006). Statis-

tical Methods for Automatic Diacritization of Arabic text. Proceedings 18th National computer Conference NCC'18، Riyadh، March 26-29.

29. F. A. H. Al-Otaibi. (2001). Speaker-dependant continuous Arabic speech recognition. M.Sc. Thesis، King Saud University.

30. M. Alghamdi، M. Elshafei، H. Almuhtaseb. (2009). Arabic Broadcast News Transcription System. International Journal of Speech Technology، 10(4)، pp. 1572-8110.

31. H. Soltau، G. Saon، B. Kingsbury، J. Kuo، L. Mangu، D. Povey، G. Zweig. (2007). The IBM 2006 GALE Arabic ASR system. ICASSP (pp. IV-349 - IV-352).

32. K. Kirchhofl، J. Bilmes، S. Das، N. Duta، M. Egan، G. Ji، F. He، J. Henderson، D. Liu، M. Noamany، P. Schoner، R. Schwartz، D. Vergyri. (2003). Novel approaches to Arabic speech recognition: report from the 2002 John-Hopkins summer workshop. ICASSP (pp. I-344–I-347).

33. H. Hyassat، R.A. Zitar. (2006). Arabic speech recognition using SPHINX engine. International Journal of Speech Technology، 9(3-4)، pp. 1381-2416.

34. R.A. Haraty، O.E. Ariss. (2007). CASRA+: A Colloquial Arabic Speech Recognition Application. American Journal of Applied Sciences 4(1)، pp. 23-32.

35. E. Mourtaga، M. Abdallah، A. Sharieh، S. Serahn. (2005). Quranic Based Speaker-Dependent Recognition Using Triphone/HMM Model. Advanced in Modeling Series B: Signal Processing and Pattern Recognition، 48(5)، pp. 43-58.

36. E. Mourtaga، M. Abdallah، A. Sharieh، S. Serahn. (2007). Speaker-Independent Quranic Recognizer Based on Maximum Likelihood Linear Regression. PWASET، Vol. 20، pp. 376-382.

37. H. Tabbal W. Al-Falou، B. Monla. (2007). Analysis and Implementation of an Automated Delimiter of "Quranic" Verses in Audio Files using Speech Recognition Techniques. Chapter of the Book "Robust Speech Recognition and Understanding"، edited by: Michael Grimm and Kristian Kroschel. pp.460.

38. http://www.j-qaf.net

39. A.M. Aziz. (2009). Developing the j-QAF e-Learning Application for Children on Islam Obligatory Duties (Fardhu Ain) under the Topic 'Ibadat'. Master's thesis، University Utara Malaysia.

40. T. Mssraty، Q. Faryadi. (2012). Teaching the Qur'anic Recitation with Harakatt: A multimedia-based Interactive Learning Method. International Journal of Scientific & Engineering Research، Volume 3، Issue 8.

41. Z. Razak، N.J. Ibrahim، et al. 2008. Quranic Verse Recitation Recognition Module for Support in j-QAF Learning: A Review. IJCSNS، 8(8)، pp. 207-216.

42. M. Zarif، N. Mohamad، B. Bakar. (2014). Assessing Quranic Reading Proficiency in the j-QAF Programme. International Education Studies Journal، Vol. 7، No. 6; Published by Canadian Center of Science and Education.

43. http://hafss.net/

44. S. Abdou، S. Hamid، M. Rashwan، A. Samir، O. AbdelHamid، M. Shahin، W. Nazih، (2006). Computer Aided Pronunciation

Learning System Using Speech Recognition Techniques. Proceedings of the INTERSPEECH 2006-ICSLP، Pittsburgh، PA، USA.

45. A. Samir، S. M. Abdou، A. H. Khalil، M. Rashwan. (2007). Enhancing usability of CAPL system for Qur'an recitation learning. Proceedings of the INTERSPEECH 2007 - ICSLP، Antwerp، Belgium.

46. S. M. Abdou، M. Rashwan. (2014). A Computer Aided Pronunciation Learning System For Teaching The Holy Quran Recitation Rules. Proceedings of the ACS/IEEE AICCSA 2014، 10-13 November، Doha، Qatar.

٤٧. يحيى الحاج، عماد الصغير، منصور الغامدي، محمد الكنهل، عبدالله الأنصاري. (٢٠١٢ م). التعليم الآلي للقرآن الكريم. تقرير فني نهائي، مدينة الملك عبدالعزيز للعلوم والتقنية، الرياض، السعودية.

48. M. AlGhamdi، Y.O.M. Elhadj، M. AlKanhal. (2007). A manual system to segment and transcribe Arabic Speech. Proceedings of IEEE ICSPC 2007، Dubai، UAE، pp. 233–236، ISBN 1-4244-1236-6.

٤٩. يحيى الحاج، منصور الغامدي، محمد الكنهل، عبدالله الأنصاري. (٢٠٠٩ م). ذخيرة صوتية لجزء من القرآن الكريم. ندوة القرآن الكريم والتقنيات المعاصرة "تقنية المعلومات"، المدينة المنورة –المملكة العربية السعودية، ١٣- ١٥ أكتوبر.

50. Y.O.M. Elhadj، M. Alghamdi، M. Alkanhal. (2013). Phoneme-Based Recognizer to Assist Reading the Holy Quran. Recent Advances in Intelligent Informatics، Advances in Intelligent Systems and Computing، Vol. 235، pp 141-152، Springer.

51. Y.O.M. Elhadj, M. Alghamdi, M. Alkanhal. (2013). Approach for Recognizing Allophonic Sounds of the Classical Arabic Based on Quran Recitations. Theory and Practice of Natural Computing, Lecture Notes in Computer Science, Vol. 8273, pp 57-67, Springer.

52. Y.O.M. Elhadj. (2010). E-Halagat: an E-Learning System for Teaching the Holy Quran. Turkish Online Journal of Educational Technology (TOJET), Vol. 9, No 1, pp: 54-61, (ISSN: 2146-7242).

٥٣. يحيــى الحــاج. (٢٠١٠ م). التحفيــظ الآلي للقــرآن الكريم. مجلــة اتصالات جمعية الحاسبات العربية، المجلد ٣، العدد ١.

٥٤. عمــاد الصغــير، عبــد الله الأنصــاري، احمــد خرصــي، يوســف العوهلــي. (٢٠٠٩م). محــرك بحــث المتشابــه اللفظي في القــرآن. ندوة القــرآن الكريم والتقنيــات المعاصــرة "تقنيــة المعلومات"، المدينــة المنورة – المملكــة العربية السعودية، ١٣–١٥ أكتوبر.

الخاتمة

استقصى هذا الكتاب ما وصلت إليه العديد من المجالات التي يتدخل فيها الحاسوب اليوم لخدمة اللغة العربية، وإن كانت ثمة نتيجة مستخلصة في نهايته، فهي أن الحاجة لتطويرها -جميعا- ما زالت كبيرة. فللوهلة الأولى -مثلا- قد يبدو أن نظام تمثيل الكتابة العربية الحالي خال من المشكلات. ولكن ما أن نقرأ البحث الأول من هذا الكتاب حتى نرى ما أخفاه الاعتياد من مشكلات. وفي البحث الثاني، تتجلى غزارة الأبواب المفتحة للتطوير. ففي أساس محدود -وهو أساس المدونات اللغوية- ومن مكان محدود أيضا -وهو جامعة ليدز- تطل علينا بحوث متضافرة تتسع مجالاتها بتعددها وتطورها. أما البحث الثالث فقد استقصى واقع قواعد البيانات الرقمية للوثائق التراثية العربية والإسلامية وبين أن الدراسات في هذا المضمار -حاليا- تنشط على قدم وساق ولكن أكثرها لا يفي بحاجات الحوسبة بعد كما أنه لا يتوفر للباحث العربي بلغته.

واستقصى البحث الرابع من الكتاب كثيرا من الأدوات التي تؤدي وظائف تشترك في احتياجها العديد من التطبيقات اللغوية العربية، وذكر مراجع نظرية وعملية لها، كما أشار إلى بعض جوانب النقص فيها. ثم أُخلص البحث الخامس لجانب التشكيل الآلي للنصوص العربية، وللقواعد التعلمية الإحصائية منه تحديدا ونسب صحة أدائها من خطئها.

ثم ناقش البحثان الأخيران تطبيقين يمثلان أبرز الغايات للباحثين في حوسبة اللغة العربية؛ وهما: تطبيق الترجمة الآلية من العربية وإليها وتطبيق التعرف على قراءة القرآن الكريم منطوقا. وقد خلص الباحثان إلى مسيس الحاجة إلى بنى تحتية ومزيد أدوات وإلى تضافر الجهود المؤسسية والرسمية لدعم ذلك كله.

ونحـن نأمل –في ختام هذا الكتـاب من باكورة إصدارات مركز الملك عبدالله بـن عبدالعزيز الدولي لخدمة اللغة العربيـة– أن يتبني هذا المركز الدولي وأمثاله المبـادرات لتطويـر حوسبة اللغة العربيـة ابتداء من مستوى البنـى التحتية ومرورا بمستوى الأدوات ووصولا لمستوى التطبيقات. والله نسأل التوفيق لما يحبه ويرضاه.

فهرس الموضوعات